

A Neyman-Pearson Approach to Universal Erasure and List Decoding

Pierre Moulin

ECE Department, Beckman Inst., and Coordinated Science Lab.

University of Illinois at Urbana-Champaign

Urbana, IL, 61801, USA

Email: moulin@ifp.uiuc.edu *

February 3, 2008

Abstract

When information is to be transmitted over an unknown, possibly unreliable channel, an erasure option at the decoder is desirable. Using constant-composition random codes, we propose a generalization of Csiszár and Körner's Maximum Mutual Information decoder with erasure option for discrete memoryless channels. The new decoder is parameterized by a *weighting function* that is designed to optimize the fundamental tradeoff between undetected-error and erasure exponents for a compound class of channels. The class of weighting functions may be further enlarged to optimize a similar tradeoff for list decoders — in that case, undetected-error probability is replaced with average number of incorrect messages in the list. Explicit solutions are identified.

The optimal exponents admit simple expressions in terms of the sphere-packing exponent, at all rates below capacity. For small erasure exponents, these expressions coincide with those derived by Forney (1968) for symmetric channels, using Maximum a Posteriori decoding. Thus for those channels at least, ignorance of the channel law is inconsequential. Conditions for optimality of the Csiszár-Körner rule and of the simpler empirical-mutual-information thresholding rule are identified. The error exponents are evaluated numerically for the binary symmetric channel.

Keywords: error exponents, constant-composition codes, random codes, method of types, sphere packing, maximum mutual information decoder, universal decoding, erasures, list decoding, Neyman-Pearson hypothesis testing.

*This work was supported by NSF under grant CCF 06-35137 and by DARPA under the ITMANET program. A 5-page version of this paper was submitted to ISIT'08.

1 Introduction

Universal decoders have been studied extensively in the information theory literature as they are applicable to a variety of communication problems where the channel is partly or even completely unknown [1, 2]. In particular, the Maximum Mutual Information (MMI) decoder provides universally attainable error exponents for random constant-composition codes over discrete memoryless channels (DMCs). In some cases, incomplete knowledge of the channel law is inconsequential as the resulting error exponents are the same as those for maximum-likelihood decoders which know the channel law in effect.

It is often desirable to provide the receiver with an erasure option that can be exercised when the received data are deemed unreliable. For fixed channels, Forney [3] derived the decision rule that provides the optimal tradeoff between the erasure and undetected-error probabilities, analogously to the Neyman-Pearson problem for binary hypothesis testing. Forney used the same framework to optimize the performance of list decoders; the probability of undetected errors is then replaced by the expected number of incorrect messages on the list. The size of the list is a random variable which equals 1 with high probability when communication is reliable.

For unknown channels, the problem of decoding with erasures was considered by Csiszár and Körner [1]. They derived attainable pairs of undetected-error and erasure exponents for any DMC. Their work was later extended by Telatar and Gallager [4]. However neither [1] nor [4] indicated whether true universality is achievable, i.e., whether the exponents match Forney's exponents. Also they did not indicate whether their error exponents might be optimal in some weaker sense. The problem was recently revisited by Merhav and Feder [5], using a competitive minimax approach. The analysis of [5] yields lower bounds on a certain fraction of the optimal exponents. It is suggested in [5] that true universality might generally not be attainable, which would represent a fundamental difference with ordinary decoding.

This paper considers decoding with erasures for the compound DMC, with two goals in mind. The first is to construct a broad class of decision rules that can be optimized in an asymptotic Neyman-Pearson sense, analogously to universal hypothesis testing [6–8]. The second is to investigate the universality properties of the receiver, in particular conditions under which the exponents coincide with Forney's exponents. We first solve the problem of variable-size list decoders because it is simpler, and the solution to the ordinary problem of size-1 lists follows directly. We establish conditions under which our error exponents match Forney's exponents.

Following background material in Sec. 2, the main results are given in Secs. 3–5. We also observe that in some problems the compound DMC approach is overly rigid and pessimistic. For such problems we present in Sec. 6 a simple and flexible extension of our method based on the relative minimax principle. In Sec. 7 we apply our results to a class of Binary Symmetric Channels (BSC), which yields easily computable and insightful formulas. The paper concludes with a brief discussion in Sec. 8. The proofs of the main results are given in the appendices.

1.1 Notation

We use uppercase letters for random variables, lowercase letters for individual values, and boldface fonts for sequences. The probability mass function (p.m.f.) of a random variable $X \in \mathcal{X}$ is denoted

by $p_X = \{p_X(x), x \in \mathcal{X}\}$, the probability of a set Ω under p_X by $P_X(\Omega)$, and the expectation operator by \mathbb{E} . Entropy of a random variable X is denoted by $H(X)$, and mutual information between two random variables X and Y is denoted by $I(X; Y) = H(X) - H(X|Y)$, or by $I(p_{XY})$ when the dependency on p_{XY} should be explicit. The Kullback-Leibler divergence between two p.m.f.'s p and q is denoted by $D(p||q)$. All logarithms are in base 2. We denote by f' the derivative of a function f .

Denote by $p_{\mathbf{x}}$ the type of a sequence $\mathbf{x} \in \mathcal{X}^N$ ($p_{\mathbf{x}}$ is an empirical p.m.f. over \mathcal{X}) and by $T_{\mathbf{x}}$ the type class associated with $p_{\mathbf{x}}$, i.e., the set of all sequences of type $p_{\mathbf{x}}$. Likewise, denote by $p_{\mathbf{xy}}$ the joint type of a pair of sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^N \times \mathcal{Y}^N$ (a p.m.f. over $\mathcal{X} \times \mathcal{Y}$) and by $T_{\mathbf{xy}}$ the type class associated with $p_{\mathbf{xy}}$, i.e., the set of all sequences of type $p_{\mathbf{xy}}$. The conditional type $p_{\mathbf{y}|\mathbf{x}}$ of a pair of sequences (\mathbf{x}, \mathbf{y}) is defined as $p_{\mathbf{xy}}(x, y)/p_{\mathbf{x}}(x)$ for all $x \in \mathcal{X}$ such that $p_{\mathbf{x}}(x) > 0$. The conditional type class $T_{\mathbf{y}|\mathbf{x}}$ is the set of all sequences $\tilde{\mathbf{y}}$ such that $(\mathbf{x}, \tilde{\mathbf{y}}) \in T_{\mathbf{xy}}$. We denote by $H(\mathbf{x})$ the entropy of the p.m.f. $p_{\mathbf{x}}$ and by $H(\mathbf{y}|\mathbf{x})$ and $I(\mathbf{x}; \mathbf{y})$ the conditional entropy and the mutual information for the joint p.m.f. $p_{\mathbf{xy}}$, respectively. Recall that [1]

$$(N+1)^{-|\mathcal{X}|} 2^{NH(\mathbf{x})} \leq |T_{\mathbf{x}}| \leq 2^{NH(\mathbf{x})}, \quad (1.1)$$

$$(N+1)^{-|\mathcal{X}||\mathcal{Y}|} 2^{NH(\mathbf{y}|\mathbf{x})} \leq |T_{\mathbf{y}|\mathbf{x}}| \leq 2^{NH(\mathbf{y}|\mathbf{x})}. \quad (1.2)$$

We let \mathcal{P}_X and $\mathcal{P}_X^{[N]}$ represent the set of all p.m.f.'s and empirical p.m.f.'s, respectively, for a random variable X . Likewise, $\mathcal{P}_{Y|X}$ and $\mathcal{P}_{Y|X}^{[N]}$ denote the set of all conditional p.m.f.'s and all empirical conditional p.m.f.'s, respectively, for a random variable Y given X . The notation $f(N) \sim g(N)$ denotes asymptotic equality: $\lim_{N \rightarrow \infty} \frac{f(N)}{g(N)} = 1$. The shorthands $f(N) \doteq g(N)$ and $f(N) \dot{\leq} g(N)$ denote equality and inequality on the exponential scale: $\lim_{N \rightarrow \infty} \frac{1}{N} \ln \frac{f(N)}{g(N)} = 0$ and $\lim_{N \rightarrow \infty} \frac{1}{N} \ln \frac{f(N)}{g(N)} \leq 0$, respectively. We denote by $\mathbb{1}_{\{x \in \Omega\}}$ the indicator function of a set Ω and define $|t|^+ \triangleq \max(0, t)$ and $\exp_2(t) \triangleq 2^t$. We adopt the notational convention that the minimum of a function over an empty set is $+\infty$.

The function-ordering notation $F \preceq G$ indicates that $F(t) \leq G(t)$ for all t . Similarly, $F \succeq G$ indicates that $F(t) \geq G(t)$ for all t .

2 Decoding with Erasure and List Options

2.1 Maximum-Likelihood Decoding

In his 1968 paper [3], Forney studied the following erasure/list decoding problem. A length- N , rate- R code $\mathcal{C} = \{\mathbf{x}(m), m \in \mathcal{M}\}$ is selected, where $\mathcal{M} = \{1, 2, \dots, 2^{NR}\}$ is the message set and each codeword $\mathbf{x}(m) \in \mathcal{X}^N$. Upon selection of a message m , the corresponding $\mathbf{x}(m)$ is transmitted over a DMC $p_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$. A set of decoding regions $\mathcal{D}_m \subseteq \mathcal{Y}^N$, $m \in \mathcal{M}$, is defined, and the decoder returns $\hat{m} = g(\mathbf{y})$ if and only if $\mathbf{y} \in \mathcal{D}_m$. For ordinary decoding, $\{\mathcal{D}_m, m \in \mathcal{M}\}$ form a partition of \mathcal{Y}^N . When an erasure option is introduced, the decision space is extended to $\mathcal{M} \cup \emptyset$, where \emptyset denotes the erasure symbol. The erasure region \mathcal{D}_{\emptyset} is the complement of $\cup_{m \in \mathcal{M}} \mathcal{D}_m$ in \mathcal{Y}^N . An undetected error arises if m was transmitted but \mathbf{y} lies in the decoding region of some other

message $i \neq m$. This event is given by

$$\mathcal{E}_i = \left\{ (m, \mathbf{y}) : \mathbf{y} \in \bigcup_{i \in \mathcal{M} \setminus \{m\}} \mathcal{D}_i \right\}, \quad (2.1)$$

where the subscript i stands for “incorrect message”. Hence

$$\begin{aligned} Pr[\mathcal{E}_i] &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{\mathbf{y} \in \bigcup_{i \in \mathcal{M} \setminus \{m\}} \mathcal{D}_i} p_{Y|X}^N(\mathbf{y}|\mathbf{x}(m)) \\ &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{i \in \mathcal{M} \setminus \{m\}} \sum_{\mathbf{y} \in \mathcal{D}_i} p_{Y|X}^N(\mathbf{y}|\mathbf{x}(m)), \end{aligned} \quad (2.2)$$

where the second equality holds because the decoding regions are disjoint.

The erasure event is given by

$$\mathcal{E}_\emptyset = \{(m, \mathbf{y}) : \mathbf{y} \in \mathcal{D}_\emptyset\}$$

and has probability

$$Pr[\mathcal{E}_\emptyset] = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{\mathbf{y} \in \mathcal{D}_\emptyset} p_{Y|X}^N(\mathbf{y}|\mathbf{x}(m)). \quad (2.3)$$

The total error event is given by $\mathcal{E}_{err} = \mathcal{E}_i \cup \mathcal{E}_\emptyset$. The decoder is generally designed so that $Pr[\mathcal{E}_i] \ll Pr[\mathcal{E}_\emptyset]$, so $Pr[\mathcal{E}_{err}] \approx Pr[\mathcal{E}_\emptyset]$.

Analogously to the Neyman-Pearson problem, one wishes to design the decoding regions to obtain an optimal tradeoff between $Pr[\mathcal{E}_i]$ and $Pr[\mathcal{E}_\emptyset]$. Forney proved the following class of decision rules is optimal:

$$g_{ML}(\mathbf{y}) = \begin{cases} \hat{m} & : \text{ if } p_{Y|X}^N(\mathbf{y}|\mathbf{x}(\hat{m})) > e^{NT} \sum_{i \neq \hat{m}} p_{Y|X}^N(\mathbf{y}|\mathbf{x}(i)) \\ \emptyset & : \text{ else} \end{cases} \quad (2.4)$$

where $T \geq 0$ is a free parameter trading off $Pr[\mathcal{E}_i]$ against $Pr[\mathcal{E}_\emptyset]$. The nonnegativity constraint on T ensures that \hat{m} is uniquely defined for any given \mathbf{y} . There is no other decision rule that yields simultaneously a lower value for $Pr[\mathcal{E}_i]$ and for $Pr[\mathcal{E}_\emptyset]$.

A conceptually simple (but suboptimal) alternative to (2.4) is

$$g_{ML,2}(\mathbf{y}) = \begin{cases} \hat{m} & : \text{ if } p_{Y|X}^N(\mathbf{y}|\mathbf{x}(\hat{m})) > e^{NT} \max_{i \neq \hat{m}} p_{Y|X}^N(\mathbf{y}|\mathbf{x}(i)) \\ \emptyset & : \text{ else} \end{cases} \quad (2.5)$$

where the decision is made based on the two highest likelihood scores.

If one chooses $T < 0$, there is generally more than one value of \hat{m} that satisfies (2.4), and g_{ML} may be viewed as a list decoder that returns the list of all such \hat{m} . Denote by N_i the number of incorrect messages on the list. Since the decoding regions $\{\mathcal{D}_m, m \in \mathcal{M}\}$ overlap, the average number of incorrect messages in the list,

$$\mathbb{E}[N_i] = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{i \in \mathcal{M} \setminus \{m\}} \sum_{\mathbf{y} \in \mathcal{D}_i} p_{Y|X}^N(\mathbf{y}|\mathbf{x}(m)), \quad (2.6)$$

no longer coincides with $Pr[\mathcal{E}_i]$ in (2.2).

For the rule (2.4) applied to symmetric channels, Forney showed that the following error exponents are achievable for all Δ such that $R^{conj} \leq R + \Delta \leq C$:

$$\begin{aligned} E_i(R, \Delta) &= E_{sp}(R) + \Delta \\ E_\emptyset(R, \Delta) &= E_{sp}(R + \Delta) \end{aligned} \quad (2.7)$$

where $E_{sp}(R)$ is the sphere-packing exponent, and R^{conj} is the conjugate rate, defined as the rate for which the slope of $E_{sp}(\cdot)$ is the inverse of the slope at rate R :

$$E'_{sp}(R^{conj}) = \frac{1}{E'_{sp}(R)}.$$

The exponents of (2.7) are achieved using independent and identically distributed (i.i.d.) codes.

2.2 Universal Decoding

When the channel law $p_{Y|X}$ is unknown, maximum-likelihood decoding cannot be used. For constant-composition codes with type p_X , the MMI decoder takes the form

$$g_{MMI}(\mathbf{y}) = \arg \max_{i \in \mathcal{M}} I(\mathbf{x}(i); \mathbf{y}). \quad (2.8)$$

Csiszár and Körner [1, p. 174—178] extended the MMI decoder to include an erasure option, using the following decision rule:

$$g_{\lambda, \Delta}(\mathbf{y}) = \begin{cases} \hat{m} & : \text{if } I(\mathbf{x}(\hat{m}); \mathbf{y}) > R + \Delta + \lambda \max_{i \neq \hat{m}} |I(\mathbf{x}(i); \mathbf{y}) - R|^+ \\ \emptyset & : \text{else} \end{cases} \quad (2.9)$$

where $\Delta \geq 0$ and $\lambda > 1$. They derived the following error exponents for the resulting undetected-error and erasure events:

$$\{E_{r, \lambda}(R, p_X, p_{Y|X}) + \Delta, E_{r, 1/\lambda}(R + \Delta, p_X, p_{Y|X})\}, \quad \forall p_{Y|X}$$

where

$$E_{r, \lambda}(R, p_X, p_{Y|X}) = \min_{\tilde{p}_{Y|X}} \{D(\tilde{p}_{Y|X} \| p_{Y|X} | p_X) + \lambda |I(p_X, \tilde{p}_{Y|X}) - R|^+\}.$$

While Δ and λ are tradeoff parameters, they did not mention whether the decision rule (2.9) satisfies any Neyman-Pearson type optimality criterion.

A different approach was recently proposed by Merhav and Feder [5]. They raised the possibility that the achievable pairs of undetected-error and erasure exponents might be smaller than in the known-channel case and proposed a decision rule based on the competitive minimax principle. This rule is parameterized by a scalar parameter $0 \leq \xi \leq 1$ which represents a fraction of the optimal exponents (for the known-channel case) that their decoding procedure is guaranteed to achieve. Decoding involves explicit maximization of a cost function over the compound DMC family, analogously to a Generalized Likelihood Ratio Test (GLRT). The rule coincides with the GLRT when $\xi = 0$, but the choice of ξ can be optimized. They conjectured that the highest achievable ξ is lower than 1 in general, and derived a computable lower bound on that value.

3 \mathcal{F} -MMI Class of Decoders

3.1 Decoding Rule

Assume that random constant-composition codes with type p_X are used, and that the DMC $p_{Y|X}$ belongs to a *connected subset* \mathcal{W} of $\mathcal{P}_{Y|X}$. The decoder knows \mathcal{W} but not which $p_{Y|X}$ is in effect.

Analogously to (2.9), our proposed decoding rule is a test based on the empirical mutual informations for the two highest-scoring messages. Let \mathcal{F} be the class of continuous, nondecreasing functions $F : [-R, H(p_X) - R] \rightarrow \mathbb{R}$. The decision rule indexed by $F \in \mathcal{F}$ takes the form

$$g_F(\mathbf{y}) = \begin{cases} \hat{m} & : \text{ if } I(\mathbf{x}(\hat{m}); \mathbf{y}) > R + \max_{i \neq \hat{m}} F(I(\mathbf{x}(i); \mathbf{y}) - R) \\ \emptyset & : \text{ else.} \end{cases} \quad (3.1)$$

Given a candidate message \hat{m} , the function F weighs the score of the best competing codeword. Since $0 \leq I(\mathbf{x}(i); \mathbf{y}) \leq H(p_X)$, all values of $F(t)$ outside the range $[-R, H(p_X) - R]$ are equivalent in terms of the decision rule (3.1).

The choice $F(t) = t$ results in the MMI decoding rule (2.8), and

$$F(t) = \Delta + \lambda |t|^+ \quad (3.2)$$

(two-parameter family of functions) results in the Csiszár-Körner rule (2.9).

One may further require that $F(t) \geq t$ to guarantee that $\hat{m} = \arg \max_i I(\mathbf{x}(i); \mathbf{y})$, as can be verified by direct substitution into (3.1). In this case, the decision is whether the decoder should output the highest-scoring message or output an erasure decision.

When the restriction $F(t) \geq t$ is not imposed, the decision rule (3.1) is ambiguous because more than one \hat{m} could satisfy the inequality in (3.1). Then (3.1) may be viewed as a list decoder that returns the list of all such \hat{m} , similarly to (2.4).

The Csiszár-Körner decision rule parameterized by F in (3.2) is nonambiguous for $\lambda \geq 1$. Note there is an error in Theorem 5.11 and Corollary 5.11A of [1, p. 175], where the condition $\lambda > 0$ should be replaced with $\lambda \geq 1$ [9].

In the limit as $\lambda \downarrow 0$, (3.2) leads to the simple decoder that lists all messages whose empirical mutual information score exceeds $R + \Delta$. If a list decoder is not desired, a simple variation on (3.2) when $0 < \lambda < 1$ is

$$F(t) = \begin{cases} \Delta + \lambda |t|^+ & : t \leq \frac{\Delta}{1-\lambda} \\ t & : \text{ else.} \end{cases}$$

It is also worth noting that the function $F(t) = \Delta + t$ may be thought of as an empirical version of Forney's suboptimal decoding rule (2.5), with $T = \Delta$. Indeed, using the identity $I(\mathbf{x}; \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x})$ and viewing the negative empirical equivocation

$$-H(\mathbf{y}|\mathbf{x}) = \sum_{x,y} p_{\mathbf{xy}}(x, y) \ln p_{\mathbf{y}|\mathbf{x}}(y|x)$$

as an empirical version of the normalized loglikelihood

$$\frac{1}{N} \ln p_{Y|X}^N(\mathbf{y}|\mathbf{x}) = \sum_{x,y} p_{\mathbf{xy}}(x, y) \ln p_{Y|X}(y|x),$$

we may rewrite (2.5) and (3.1) respectively as

$$g_{ML,2}(\mathbf{y}) = \begin{cases} \hat{m} & : \text{ if } \frac{1}{N} \ln p_{Y|X}^N(\mathbf{y}|\mathbf{x}(\hat{m})) > T + \max_{i \neq \hat{m}} \frac{1}{N} \ln p_{Y|X}^N(\mathbf{y}|\mathbf{x}(i)) \\ \emptyset & : \text{ else} \end{cases} \quad (3.3)$$

and

$$g_F(\mathbf{y}) = \begin{cases} \hat{m} & : \text{ if } -H(\mathbf{y}|\mathbf{x}(\hat{m})) > \Delta + \max_{i \neq \hat{m}} [-H(\mathbf{y}|\mathbf{x}(i))] \\ \emptyset & : \text{ else.} \end{cases} \quad (3.4)$$

While this observation does not imply $F(t) = \Delta + t$ is an optimal choice for F , one might intuitively expect optimality in some regime.

3.2 Error Exponents

For a random-coding strategy using constant-composition codes with type p_X , the expected number of incorrect messages on the list, $\mathbb{E}[N_i]$, and the erasure probability, $Pr[\mathcal{E}_\emptyset]$, may be viewed as functions of R , p_X , $p_{Y|X}$, and F . A pair $\{E_i(R, p_X, p_{Y|X}, F), E_\emptyset(R, p_X, p_{Y|X}, F)\}$ of incorrect-message and erasure exponents is said to be universally attainable for such codes over \mathscr{W} if the expected number of incorrect messages on the list and the erasure probability satisfy

$$\mathbb{E}[N_i] \leq \exp_2 \{-N[E_i(R, p_X, p_{Y|X}, F) - \epsilon]\}, \quad (3.5)$$

$$Pr[\mathcal{E}_\emptyset] \leq \exp_2 \{-N[E_\emptyset(R, p_X, p_{Y|X}, F) - \epsilon]\}, \quad \forall p_{Y|X} \in \mathscr{W}, \quad (3.6)$$

for any $\epsilon > 0$ and N greater than some $N_0(\epsilon)$. The worst-case exponents (over all $p_{Y|X} \in \mathscr{W}$) are denoted by

$$E_i(R, p_X, \mathscr{W}, F) \triangleq \min_{p_{Y|X} \in \mathscr{W}} E_i(R, p_X, p_{Y|X}, F), \quad (3.7)$$

$$E_\emptyset(R, p_X, \mathscr{W}, F) \triangleq \min_{p_{Y|X} \in \mathscr{W}} E_\emptyset(R, p_X, p_{Y|X}, F). \quad (3.8)$$

Our problem is to maximize the erasure exponent $E_\emptyset(R, p_X, \mathscr{W}, F)$ subject to the constraint that the incorrect-message exponent $E_i(R, p_X, \mathscr{W}, F)$ is at least equal to some prescribed value α . This is an asymptotic Neyman-Pearson problem. We shall focus on the regime of practical interest where erasures are more acceptable than undetected errors:

$$E_\emptyset(R, p_X, \mathscr{W}, F) \leq E_i(R, p_X, \mathscr{W}, F).$$

We emphasize that asymptotic Neyman-Pearson optimality of the decision rule holds only in a restricted sense, namely, with respect to the \mathcal{F} -MMI class (3.1).

Specifically, given R and \mathscr{W} , we seek the solution to the constrained optimization problem

$$E_\emptyset^*(R, \mathscr{W}, \alpha) \triangleq \max_{p_X} \max_{F \in \mathcal{F}(R, p_X, \mathscr{W}, \alpha)} \min_{p_{Y|X} \in \mathscr{W}} E_\emptyset(R, p_X, p_{Y|X}, F) \quad (3.9)$$

where $\mathcal{F}(R, p_X, \mathscr{W}, \alpha)$ is the set of functions F that satisfy

$$\min_{p_{Y|X} \in \mathscr{W}} E_i(R, p_X, p_{Y|X}, F) \geq \alpha \quad (3.10)$$

as well as the continuity and monotonicity conditions mentioned above (3.1).

If we were able to choose F as a function of $p_{Y|X}$, we would do at least as well as in (3.9) and achieve the erasure exponent

$$\begin{aligned} E_{\emptyset}^{**}(R, \mathcal{W}, \alpha) &\triangleq \max_{p_X} \min_{p_{Y|X} \in \mathcal{W}} \max_{F \in \mathcal{F}(R, p_X, p_{Y|X}, \alpha)} E_{\emptyset}(R, p_X, p_{Y|X}, F) \\ &\geq E_{\emptyset}^*(R, \mathcal{W}, \alpha). \end{aligned} \quad (3.11)$$

We shall be particularly interested in characterizing (R, \mathcal{W}, α) for which the decoder incurs no penalty for not knowing $p_{Y|X}$, i.e.,

- equality holds in (3.11), and
- the optimal exponents $E_i(R, p_X, p_{Y|X}, F)$ and $E_{\emptyset}(R, p_X, p_{Y|X}, F)$ in (3.10) and (3.9) coincide with Forney's exponents in (2.7) for all $p_{Y|X} \in \mathcal{W}$,

the second property being stronger than the first one.

3.3 Basic Properties of F

To simplify the derivations, it is convenient to slightly strengthen the requirement that F be nondecreasing, and work with strictly increasing functions F instead. Then the maxima over F in (3.9) and (3.11) are replaced with suprema, but of course their value remains the same.

To each monotonically increasing function F corresponds an inverse F^{-1} , such that

$$F(t) = u \quad \Leftrightarrow \quad F^{-1}(u) = t.$$

Elementary properties satisfied by the inverse function include:

- (P1) F^{-1} is continuous and increasing over its range.
- (P2) If $F \preceq G$, then $F^{-1} \succeq G^{-1}$.
- (P3) $G(t) = F(t) + \Delta \quad \Leftrightarrow \quad G^{-1}(t) = F^{-1}(t - \Delta)$.
- (P4) $\frac{dF(t)}{dt} = 1 \Big/ \frac{dF^{-1}(t)}{dt}$.
- (P5) If F is convex, then F^{-1} is concave.
- (P6) The domain of F^{-1} is the range of F , and vice-versa.

Now for any F such that $F(H(p_X) - R) \geq 0$, define the scalar

$$t_F \triangleq \sup \{t : F(t) = |F(t')| \quad \forall t' \leq t \leq H(p_X) - R\} \quad (3.12)$$

which may depend on R and p_X via the difference $H(p_X) - R$. From this definition we have the following properties:

- $|F(t)|^+$ is constant for all $t \leq t_F$;
- $F(t) \geq 0$ for $t \geq t_F$.

We have $t_F = 0$ if $F(t)$ is chosen as in (4.5), or if $F(t) = \Delta + \lambda|t|^+$. If F has a zero-crossing, t_F is that zero-crossing. For instance, $t_F = -\Delta/\lambda$ if $F(t) = \Delta + \lambda t$. Or $t_F = \min\{\Delta, H(p_X) - R\}$ if $F(t) = a|t - \Delta|^+$. The supremum in (3.12) always exists.

4 Random-Coding and Sphere-Packing Exponents

The sphere packing exponent for channel $p_{Y|X}$ is defined as

$$E_{sp}(R, p_X, p_{Y|X}) \triangleq \min_{\tilde{p}_{Y|X} : I(p_X, \tilde{p}_{Y|X}) \leq R} D(\tilde{p}_{Y|X} \| p_{Y|X} | p_X) \quad (4.1)$$

and as ∞ if the minimization above is over an empty set. The function $E_{sp}(R, p_X, p_{Y|X})$ is convex, nonincreasing, and differentiable in R , and continuous in $p_{Y|X}$.

The sphere packing exponent for class \mathcal{W} is defined as

$$E_{sp}(R, p_X, \mathcal{W}) \triangleq \min_{p_{Y|X} \in \mathcal{W}} E_{sp}(R, p_X, p_{Y|X}). \quad (4.2)$$

The function $E_{sp}(R, p_X, \mathcal{W})$ is differentiable in R because \mathcal{W} is a connected set. In some cases, $E_{sp}(R, p_X, \mathcal{W})$ is convex in R , e.g., when the same $p_{Y|X}$ achieves the minimum in (4.2) at all rates. Denote by $R_\infty(p_X, \mathcal{W})$ the infimum of the rates R such that $E_{sp}(R, p_X, \mathcal{W}) < \infty$, and by $I(p_X, \mathcal{W}) = \min_{p_{Y|X} \in \mathcal{W}} I(p_X, p_{Y|X})$ the supremum of R such that $E_{sp}(R, p_X, \mathcal{W}) > 0$.

The *modified random coding exponent* for channel $p_{Y|X}$ and for class \mathcal{W} are respectively defined as

$$E_{r,F}(R, p_X, p_{Y|X}) \triangleq \min_{\tilde{p}_{Y|X}} [D(\tilde{p}_{Y|X} \| p_{Y|X} | p_X) + F(I(p_X, \tilde{p}_{Y|X}) - R)] \quad (4.3)$$

and

$$E_{r,F}(R, p_X, \mathcal{W}) \triangleq \min_{p_{Y|X} \in \mathcal{W}} E_{r,F}(R, p_X, p_{Y|X}). \quad (4.4)$$

When $F(t) = |t|^+$, (4.3) is just the usual random coding exponent. It can be verified that (4.3) is a continuous functional of F .

Define the function

$$F_{R,p_X,\mathcal{W}}(t) \triangleq E_{sp}(R, p_X, \mathcal{W}) - E_{sp}(R+t, p_X, \mathcal{W}) \quad (4.5)$$

which is depicted in Fig. 1 for a BSC example to be analyzed in Sec. 7. This function is increasing for $R_\infty(p_X, \mathcal{W}) - R \leq t \leq I(p_X, \mathcal{W}) - R$ and satisfies the following properties:

$$\begin{aligned} F_{R,p_X,\mathcal{W}}(0) &= 0 \\ F'_{R,p_X,\mathcal{W}}(t) &= -E'_{sp}(R+t, p_X, \mathcal{W}) \\ E_{sp}(R', p_X, \mathcal{W}) + F_{R,p_X,\mathcal{W}}(R' - R) &\equiv E_{sp}(R, p_X, \mathcal{W}). \end{aligned} \quad (4.6)$$

If $E_{sp}(R, p_X, \mathcal{W})$ is convex in R , then $F_{R,p_X,\mathcal{W}}(t)$ is concave in t .

Proposition 4.1 *The modified random coding exponent $E_{r,F}(R, p_X, p_{Y|X})$ satisfies the following properties.*

- (i) $E_{r,F}(R, p_X, p_{Y|X})$ is nonincreasing in R .
- (ii) If $F \prec G$, then $E_{r,F}(R, p_X, p_{Y|X}) \leq E_{r,G}(R, p_X, p_{Y|X})$.
- (iii) $E_{r,F}(R, p_X, p_{Y|X})$ is related to the sphere packing exponent as follows:

$$E_{r,F}(R, p_X, p_{Y|X}) = \min_{R'} [E_{sp}(R', p_X, p_{Y|X}) + F(R' - R)]. \quad (4.7)$$

- (iv) The above properties hold with \mathcal{W} in place of $p_{Y|X}$ in the arguments of the functions $E_{r,F}$ and E_{sp} .

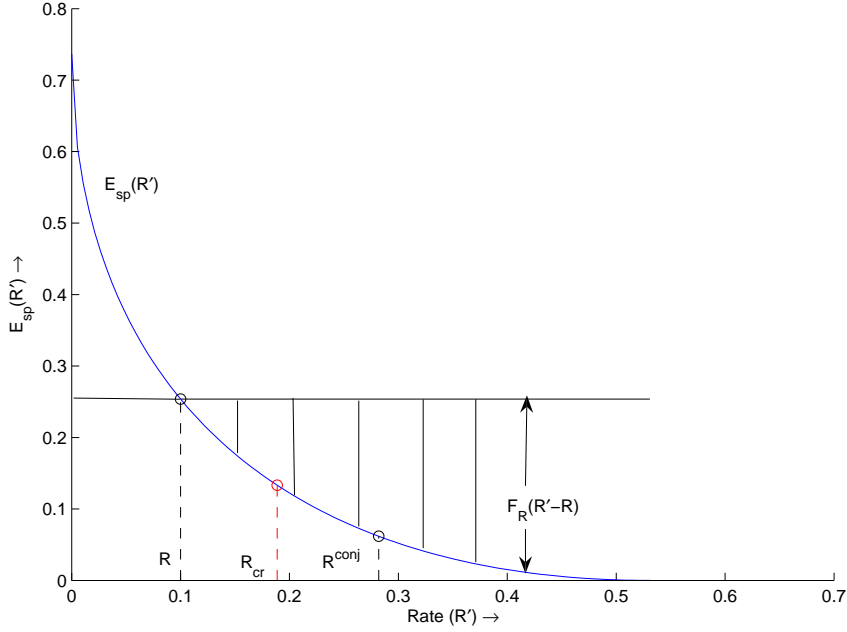


Figure 1: Function $F_{R,p_X,\mathcal{W}}(\cdot - R)$ when $R = 0.1$, \mathcal{W} is the family of BSC's with crossover probability $\rho \leq 0.1$ (capacity $C(\rho) \geq 0.53$), and p_X is the uniform p.m.f. over $\{0, 1\}$.

The proof of these properties is given in the appendix. Part (iii) is a variation on Lemma 5.4 and its corollary in [1, p. 168]. Also note that while $E_{r,F}(R, p_X, p_{Y|X})$ is convex in R for some choices of F , including (3.2), that property does not extend to arbitrary F .

Proposition 4.2 *The incorrect-message and erasure exponents under the decision rule (3.1) are respectively given by*

$$E_i(R, p_X, p_{Y|X}, F) = E_{r,F}(R, p_X, p_{Y|X}), \quad (4.8)$$

$$E_\emptyset(R, p_X, p_{Y|X}, F) = E_{r,|F^{-1}|^+}(R, p_X, p_{Y|X}). \quad (4.9)$$

Proof: see appendix.

If $F(t) = |t|^+$, then $|F^{-1}(t)|^+ = |t|^+$, and both (4.8) and (4.9) reduce to the ordinary random-coding exponent.

If the channel is not reliable enough in the sense that $I(p_X, p_{Y|X}) \leq R + F(0)$, then $F^{-1}(I(p_X, p_{Y|X}) - R) \leq F^{-1}(F(0)) = 0$, and from (4.9) we obtain $E_\emptyset(R, p_X, p_{Y|X}, F) = 0$ because the minimizing $\tilde{p}_{Y|X}$ in the expression for $E_{r,|F^{-1}|^+}$ is equal to $p_{Y|X}$. Checking (3.1), a heuristic explanation for the zero erasure exponent is that $I(\mathbf{x}(m); \mathbf{y}) \approx I(p_X, p_{Y|X})$ with high probability when m is the transmitted message, and $\max_{i \neq m} I(\mathbf{x}(m); \mathbf{y}) \approx R$ (obtained using (B.4) with $\nu = R$ and the union bound).

5 \mathcal{W} -Optimal Choice of F

In this section, we view the incorrect-message and erasure exponents as functionals of $|F|^+$ and examine optimal tradeoffs between them. It is instructive to first consider the one-parameter family

$$F(t) \equiv \Delta \geq 0, \quad (5.1)$$

which corresponds to a thresholding rule in (3.1). Using (4.7) it is easily verified that

$$E_i(R, p_X, p_{Y|X}, F) = E_{r,F}(R, p_X, p_{Y|X}) \equiv \Delta, \quad (5.2)$$

$$E_\emptyset(R, p_X, p_{Y|X}, F) \equiv E_{sp}(R + \Delta, p_X, p_{Y|X}). \quad (5.3)$$

Two extreme choices for Δ are 0 and $I(p_X, \mathcal{W}) - R$ because in each case one error exponent is zero and the other one is positive. One would expect that better tradeoffs can be achieved using a broader class of functions F , though.

Recalling (3.9) (3.10) and using (4.8) and (4.9), we seek the solution to the following two asymptotic Neyman-Pearson optimization problems. For list decoding, find

$$E_\emptyset^L(R, \mathcal{W}, \alpha) = \max_{p_X} E_\emptyset^L(R, p_X, \mathcal{W}, \alpha) \quad (5.4)$$

where the cost function

$$E_\emptyset^L(R, p_X, \mathcal{W}, \alpha) = \sup_{F \in \mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)} E_{r, |F^{-1}|^+}(R, p_X, \mathcal{W}). \quad (5.5)$$

where the feasible set $\mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$ is the set of continuous, increasing functions F that yield an incorrect-message exponent at least equal to α :

$$E_i(R, p_X, \mathcal{W}, F) = E_{r,F}(R, p_X, \mathcal{W}) \geq \alpha. \quad (5.6)$$

For classical decoding (list size ≤ 1), find

$$E_\emptyset(R, \mathcal{W}, \alpha) = \max_{p_X} E_\emptyset(R, p_X, \mathcal{W}, \alpha) \quad (5.7)$$

where

$$E_\emptyset(R, p_X, \mathcal{W}, \alpha) = \sup_{F \in \mathcal{F}(R, p_X, \mathcal{W}, \alpha)} E_{r, |F^{-1}|^+}(R, p_X, \mathcal{W}) \quad (5.8)$$

and $\mathcal{F}(R, p_X, \mathcal{W}, \alpha)$ is the subset of functions in $\mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$ that satisfy $F(t) \geq t$, i.e., the decoder outputs at most one message.

Since $\mathcal{F}(R, p_X, \mathcal{W}, \alpha) \subset \mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$, we have

$$E_\emptyset(R, p_X, \mathcal{W}, \alpha) \leq E_\emptyset^L(R, p_X, \mathcal{W}, \alpha).$$

The sequel of this paper focuses on the class of variable-size list decoders associated with \mathcal{F}^L because the error exponent tradeoffs are at least as good as those associated with \mathcal{F} , and the corresponding error exponents take a more concise form.

Define the critical rate $R_{cr}(p_X, \mathcal{W})$ as the rate at which the derivative $E'_{sp}(\cdot, p_X, \mathcal{W}) = -1$, and

$$\Delta \triangleq \alpha - E_{sp}(R, p_X, \mathcal{W}). \quad (5.9)$$

Two rates R_1 and R_2 are said to be conjugate given p_X and \mathcal{W} if the corresponding slopes of $E_{sp}(\cdot, p_X, \mathcal{W})$ are inverse of each other:

$$E'_{sp}(R_1, p_X, \mathcal{W}) = \frac{1}{E'_{sp}(R_2, p_X, \mathcal{W})}. \quad (5.10)$$

The difference d between two conjugate rates uniquely specifies them. We denote the smaller one by $R_1(d)$ and the larger one by $R_2(d)$, irrespective of the sign of d . Hence $R_1(d) \leq R_{cr}(p_X, \mathcal{W}) \leq R_2(d)$, with equality when $d = 0$. We also denote by $R^{conj}(p_X, \mathcal{W})$ the conjugate rate of R , as defined by (5.10). The conjugate rate always exists when R is below the critical rate $R_{cr}(p_X, \mathcal{W})$. If R is above the critical rate and sufficiently large, $R^{conj}(p_X, \mathcal{W})$ may not exist. Instead of treating this case separately, we note that this case will be irrelevant because the conjugate rate always appears via the expression $\max\{R, R^{conj}(p_X, \mathcal{W})\}$ which is equal to R if $R > R_{cr}(p_X, \mathcal{W})$ and therefore this expression is always well defined.

The proofs of Prop. 5.1 and 5.3 and Lemma 5.4 below may be found in the appendix; recall $F_{R, p_X, \mathcal{W}}(t)$ was defined in (4.5). The proof of Prop. 5.5 parallels that of Prop. 5.1(ii) and is therefore omitted.

Proposition 5.1 *The suprema in (5.5) and (5.8) are respectively achieved by*

$$F^{L*}(t) = F_{R, p_X, \mathcal{W}}(t) + \Delta \quad (5.11)$$

$$= \alpha - E_{sp}(R + t, p_X, \mathcal{W}) \quad (5.12)$$

and

$$F^*(t) = \max(t, F^{L*}(t)). \quad (5.13)$$

The resulting incorrect-message exponent is given by $E_i(R, p_X, \mathcal{W}) = \alpha$. The optimal solution is nonunique. In particular, for $t \leq 0$, one can replace $F^{L}(t)$ by the constant $F^{L*}(0)$ without effect on the error exponents.*

The proof of Prop. 5.1 is quite simple and can be separated from the calculation of the error exponents. The main idea is as follows. If $G \preceq F$, we have $E_{r, G}(R, p_X, \mathcal{W}) \leq E_{r, F}(R, p_X, \mathcal{W})$. Since $G^{-1} \succeq F^{-1}$, we also have $E_{r, G^{-1}}(R, p_X, \mathcal{W}) \geq E_{r, F^{-1}}(R, p_X, \mathcal{W})$. Therefore we seek $F^* \in \mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$ such that $F^* \preceq F$ for all $F \in \mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$. Such F^* , assuming it exists, necessarily achieves $E_\emptyset^L(R, p_X, \mathcal{W}, \alpha)$. The same procedure applies to $\mathcal{F}(R, p_X, \mathcal{W}, \alpha)$.

Corollary 5.2 *If $R \geq I(p_X, \mathcal{W})$, the thresholding rule $F^{L*}(t) \equiv \Delta$ of (5.1) is optimal, and the optimal error exponents are $E_i(R, p_X, \mathcal{W}, \alpha) = \Delta$ and $E_\emptyset^L(R, p_X, \mathcal{W}, \alpha) = 0$.*

Proof. Since $R \geq I(p_X, \mathcal{W})$, we have $E_{sp}(R, p_X, \mathcal{W}) = 0$. Hence from (4.5), $F_{R, p_X, \mathcal{W}}(t) \equiv 0$ for all $t \geq 0$. Substituting into (5.11) proves the optimality of the thresholding rule (5.1). The corresponding error exponents are obtained by minimizing (5.2) and (5.3) over $p_{Y|X} \in \mathcal{W}$. \square

The case $R < I(p_X, \mathcal{W})$ is addressed next.

Proposition 5.3 *If $E_{sp}(R, p_X, \mathcal{W})$ is convex in R , the optimal incorrect-message and erasure exponents are related as follows.*

(i) *For $|R^{conj}(p_X, \mathcal{W}) - R|^+ \leq \Delta \leq I(p_X, \mathcal{W}) - R$, we have*

$$\begin{aligned} E_i(R, p_X, \mathcal{W}, \alpha) &= E_{sp}(R, p_X, \mathcal{W}) + \Delta = \alpha \\ E_\emptyset^L(R, p_X, \mathcal{W}, \alpha) &= E_{sp}(R + \Delta, p_X, \mathcal{W}). \end{aligned} \quad (5.14)$$

(ii) *The above exponents are also achieved using the penalty function $F(t) = \Delta + \lambda|t|^+$ with*

$$-E'_{sp}(R, p_X, \mathcal{W}) \leq \lambda \leq \frac{1}{-E'_{sp}(R + \Delta, p_X, \mathcal{W})}. \quad (5.15)$$

(iii) *If $R \leq R_{cr}(p_X, \mathcal{W})$ and $0 \leq \Delta \leq R^{conj}(p_X, \mathcal{W}) - R$, we have*

$$\begin{aligned} E_i(R, p_X, \mathcal{W}, \alpha) &= E_{sp}(R, p_X, \mathcal{W}) + \Delta = \alpha \\ E_\emptyset^L(R, p_X, \mathcal{W}, \alpha) &= E_{sp}(R_2(\Delta), p_X, \mathcal{W}) + F_{R, p_X, \mathcal{W}}^{-1}(R_1(\Delta) - R). \end{aligned} \quad (5.16)$$

(iv) *If $R \leq R_{cr}(p_X, \mathcal{W})$ and $R_\infty(p_X, \mathcal{W}) - R \leq \Delta \leq 0$, we have*

$$\begin{aligned} E_i(R, p_X, \mathcal{W}, \alpha) &= E_{sp}(R, p_X, \mathcal{W}) + \Delta = \alpha \\ E_\emptyset^L(R, p_X, \mathcal{W}, \alpha) &= E_{sp}(R_1(\Delta), p_X, \mathcal{W}) + F_{R, p_X, \mathcal{W}}^{-1}(R_2(\Delta) - R). \end{aligned} \quad (5.17)$$

Part (ii) of the proposition implies that not only is the optimal F nonunique under the combinations of $(R, p_X, \mathcal{W}, \alpha)$ of Part (i), but also the Csiszár-Körner rule (2.9) is optimal *for any* (Δ, λ) in a certain range of values.

Also, while Prop. 5.3 provides simple expressions for the worst-case error exponents over \mathcal{W} , the exponents for any specific channel $p_{Y|X} \in \mathcal{W}$ are obtained by substituting the function (5.12) and its inverse, respectively, into the minimization problem of (4.7). This problem does generally not admit a simple expression.

This leads us back to the question asked at the end of Sec. 2, namely, when does the decoder pay no penalty for not knowing $p_{Y|X}$? Defining

$$\underline{E}'_{sp}(R, p_X, \mathcal{W}) \triangleq \min_{p_{Y|X} \in \mathcal{W}} E'_{sp}(R, p_X, p_{Y|X}), \quad (5.18)$$

and

$$\overline{R}^{conj}(p_X, \mathcal{W}) \triangleq \max_{p_{Y|X} \in \mathcal{W}} R^{conj}(p_X, p_{Y|X}) \quad (5.19)$$

we have the following lemma, whose proof appears in the appendix.

Lemma 5.4

$$E'_{sp}(R, p_X, \mathcal{W}) \geq \underline{E}'_{sp}(R, p_X, \mathcal{W}), \quad (5.20)$$

$$\overline{R}^{conj}(p_X, \mathcal{W}) \geq R^{conj}(p_X, \mathcal{W}) \quad (5.21)$$

with equality if the same $p_{Y|X}$ minimizes $E_{sp}(R, p_X, p_{Y|X})$ at all rates.

Proposition 5.5 Assume that $R, p_X, \mathcal{W}, \Delta$ and λ are such that

$$|\overline{R}^{conj}(p_X, \mathcal{W}) - R|^+ \leq \Delta \leq I(p_X, \mathcal{W}) - R, \quad (5.22)$$

$$-\underline{E}'_{sp}(R, p_X, \mathcal{W}) \leq \lambda \leq \frac{1}{-\underline{E}'_{sp}(R + \Delta, p_X, \mathcal{W})}. \quad (5.23)$$

Then the pair of incorrect-message and erasure exponents

$$\{E_{sp}(R, p_X, p_{Y|X}) + \Delta, E_{sp}(R + \Delta, p_X, p_{Y|X})\} \quad (5.24)$$

is universally attainable over $p_{Y|X} \in \mathcal{W}$ using the penalty function $F(t) = \Delta + \lambda|t|^+$, and equality holds in the erasure-exponent game of (3.11).

Proof. From (5.19) and (5.22), we have

$$|R^{conj}(p_X, p_{Y|X}) - R|^+ \leq \Delta \leq I(p_X, p_{Y|X}) - R, \quad \forall p_{Y|X} \in \mathcal{W}.$$

Similarly, from (5.18) and (5.23), we have

$$-E'_{sp}(R, p_X, p_{Y|X}) \leq \lambda \leq \frac{1}{-E'_{sp}(R + \Delta, p_X, p_{Y|X})}, \quad \forall p_{Y|X} \in \mathcal{W}.$$

Then applying Prop. 5.3(ii) with the singleton $\{p_{Y|X}\}$ in place of \mathcal{W} proves the claim. \square

The set of (Δ, λ) defined by (5.22), (5.23) is smaller than that of Prop. 5.3(i) but is not empty because $\underline{E}'_{sp}(R + \Delta, p_X, \mathcal{W})$ tends to zero as Δ approaches the upper limit $I(p_X, \mathcal{W}) - R$. Thus the universal exponents in (5.24) hold at least in the small erasure-exponent regime (where $E_{sp}(R + \Delta, p_X, p_{Y|X}) \rightarrow 0$) and coincide with those derived by Forney [3, Theorem 3(a)] for symmetric channels, using MAP decoding. For symmetric channels, the same input distribution p_X is optimal at all rates.¹ Our rates are identical to his, i.e., the same optimal error exponents are achieved without knowledge of the channel.

6 Relative Minimax

When the compound class \mathcal{W} is so large that $I(p_X, \mathcal{W}) \leq R$, we have seen from Corollary 5.2 that the simple thresholding rule $F(t) \equiv \Delta$ is optimal. Even if $I(p_X, \mathcal{W}) > R$, our minimax criterion (which seeks the worst-case error exponents over the class \mathcal{W}) for designing F might be a pessimistic one. This drawback can be alleviated to some extent using a *relative minimax* principle, see [10] and references therein. Our proposed approach is to define two functionals $\alpha(p_{Y|X})$ and $\beta(p_{Y|X})$ and the *relative error exponents*

$$\begin{aligned} \Delta_\alpha E_i(R, p_X, p_{Y|X}, F) &\triangleq E_i(R, p_X, p_{Y|X}, F) - \alpha(p_{Y|X}), \\ \Delta_\beta E_\emptyset(R, p_X, p_{Y|X}, F) &\triangleq E_\emptyset(R, p_X, p_{Y|X}, F) - \beta(p_{Y|X}). \end{aligned}$$

Then solve the constrained optimization problem of (3.9) with the above functionals in place of $E_i(R, p_X, p_{Y|X}, F) - \alpha$ and $E_\emptyset(R, p_X, p_{Y|X}, F)$. It is reasonable to choose $\alpha(p_{Y|X})$ and $\beta(p_{Y|X})$

¹ Forney also studied the case $E_\emptyset(R) > E_i(R)$, which is not covered by our analysis.

large for “good channels” and small for very noisy channels. While $\alpha(p_{Y|X})$ and $\beta(p_{Y|X})$ could be the error exponents associated with some reference test, this is not a requirement. A possible choice is

$$\begin{aligned}\alpha(p_{Y|X}) &= \Delta \\ \beta(p_{Y|X}) &= E_{sp}(R + \Delta, p_X, p_{Y|X})\end{aligned}$$

which are the error exponents (5.2) and (5.3) corresponding to the thresholding rule $F(t) \equiv \Delta$. Another choice is

$$\alpha(p_{Y|X}) = E_{sp}(R, p_X, p_{Y|X}) + \Delta \quad (6.1)$$

$$\beta(p_{Y|X}) = E_{sp}(R + \Delta, p_X, p_{Y|X}) \quad (6.2)$$

which are the “ideal” Forney exponents — achievable under the assumptions of Prop. 5.5(i).

The relative minimax problem is a simple extension of the minimax problem solved earlier. Define the following functions:

$$\Delta_\alpha E_{r,F}(R, p_X, \mathcal{W}) = \min_{p_{Y|X} \in \mathcal{W}} [E_{r,F}(R, p_X, p_{Y|X}) - \alpha(p_{Y|X})], \quad (6.3)$$

$$\Delta_\alpha E_{sp}(R, p_X, \mathcal{W}) = \min_{p_{Y|X} \in \mathcal{W}} [E_{sp}(R, p_X, p_{Y|X}) - \alpha(p_{Y|X})], \quad (6.4)$$

$$F_{R,p_X,\mathcal{W},\alpha}(t) = \Delta_\alpha E_{sp}(R, p_X, \mathcal{W}) - \Delta_\alpha E_{sp}(R + t, p_X, \mathcal{W}). \quad (6.5)$$

The function $F_{R,p_X,\mathcal{W},\alpha}(t)$ of (6.5) is increasing and satisfies $F_{R,p_X,\mathcal{W},\alpha}(0) = 0$. The above functions $\Delta_\alpha E_{r,F}$ and $\Delta_\alpha E_{sp}$ satisfy the following relationship:

$$\begin{aligned}\Delta_\alpha E_{r,F}(R, p_X, \mathcal{W}) &\stackrel{(a)}{=} \min_{p_{Y|X} \in \mathcal{W}} \left\{ \min_{R'} [E_{sp}(R', p_X, p_{Y|X}) + F(R' - R)] - \alpha(p_{Y|X}) \right\} \\ &= \min_{R'} \left\{ \min_{p_{Y|X} \in \mathcal{W}} [E_{sp}(R', p_X, p_{Y|X}) - \alpha(p_{Y|X})] + F(R' - R) \right\} \\ &\stackrel{(b)}{=} \min_{R'} [\Delta_\alpha E_{sp}(R', p_X, \mathcal{W}) + F(R' - R)]\end{aligned} \quad (6.6)$$

where (a) is obtained from (4.7) and (6.3), and (b) from (6.4). Equation (6.6) is of the same form as (4.7), with $\Delta_\alpha E_{r,F}$ and $\Delta_\alpha E_{sp}$ in place of $E_{r,F} - \alpha$ and $E_{sp} - \alpha$, respectively.

Analogously to (5.4), (5.5), and (5.6), the relative minimax for variable-size decoders is given by

$$\Delta_\beta E_\emptyset^L(R, \mathcal{W}, \alpha) = \max_{p_X} \sup_{F \in \mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)} \Delta_\beta E_{r,|F^{-1}|+}(R, p_X, \mathcal{W}) \quad (6.7)$$

where the feasible set $\mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$ is the set of functions F that satisfy

$$\Delta_\alpha E_{r,F}(R, p_X, \mathcal{W}) \geq 0$$

as well as the previous continuity and monotonicity conditions. The following proposition is analogous to Prop. 5.1.

Proposition 6.1 *The supremum over F in (6.7) is achieved by*

$$\begin{aligned} F^{L*}(t) &= F_{R,p_X,\mathscr{W},\alpha}(t) - \Delta_\alpha E_{sp}(R, p_X, \mathscr{W}) \\ &= -\Delta_\alpha E_{sp}(R+t, p_X, \mathscr{W}), \end{aligned} \quad (6.8)$$

independently of the choice of β . The relative minimax is given by

$$\Delta_\beta E_\emptyset^L(R, \mathscr{W}, \alpha) = \max_{p_X} \Delta_\beta E_{r,|(F^{L*})^{-1}|}(R, p_X, \mathscr{W}).$$

Proof: The proof exploits the same monotonicity property ($E_{r,F} \leq E_{r,G}$ for $F \preceq G$) that was used to derive the optimal F in (5.12). The supremum over F is obtained by following the steps of the proof of Prop. 5.1, substituting $\Delta_\alpha E_{r,F}$, $\Delta_\alpha E_{sp}$, and $F_{R,p_X,\mathscr{W},\alpha}$ for $E_{r,F} - \alpha$, $E_{sp} - \alpha$, and $F_{R,p_X,\mathscr{W}}$, respectively. The relative minimax is obtained by substituting the optimal F into (6.7). \square

We would like to know how much influence the reference function α has on the optimal F . For the “Forney reference exponent function” α of (6.1), we obtain the optimal F from (6.8) and (6.4):

$$\begin{aligned} F^{L*}(t) &= - \min_{p_{Y|X} \in \mathscr{W}} [E_{sp}(R+t, p_X, p_{Y|X}) - \alpha(p_{Y|X})] \\ &= - \min_{p_{Y|X} \in \mathscr{W}} [E_{sp}(R+t, p_X, p_{Y|X}) - E_{sp}(R, p_X, p_{Y|X}) - \Delta] \\ &= \Delta + \max_{p_{Y|X} \in \mathscr{W}} [E_{sp}(R, p_X, p_{Y|X}) - E_{sp}(R+t, p_X, p_{Y|X})] \\ &= \Delta + \max_{p_{Y|X} \in \mathscr{W}} F_{R,p_X,p_{Y|X}}(t). \end{aligned} \quad (6.9)$$

Interestingly, the maximum above is often achieved by the *cleanest* channel in \mathscr{W} — for which $E_{sp}(R, p_X, p_{Y|X})$ is large and $E_{sp}(R+t, p_X, p_{Y|X})$ falls off rapidly as t increases. This stands in contrast to (5.12) which may be written as

$$\begin{aligned} F^{L*}(t) &= \alpha - \min_{p_{Y|X} \in \mathscr{W}} E_{sp}(R+t, p_X, p_{Y|X}) \\ &= \Delta + \min_{p_{Y|X} \in \mathscr{W}} E_{sp}(R, p_X, p_{Y|X}) - \min_{p_{Y|X} \in \mathscr{W}} E_{sp}(R+t, p_X, p_{Y|X}). \end{aligned} \quad (6.10)$$

In (6.10), the minima are achieved by the *noisiest* channel at rates R and $R+t$, respectively. Also note that $F^{L*}(t)$ from (6.9) is uniformly larger than $F^{L*}(t)$ from (6.10) and thus results in larger incorrect-message exponents.

For $R > I(p_X, \mathscr{W})$, Corollary 5.2 has shown that the minimax criterion is maximized by the thresholding rule $F(t) = \Delta$ which yields $E_i(R, p_X, p_{Y|X}) = \Delta$ and $E_\emptyset(R, p_X, p_{Y|X}) = E_{sp}(R + \Delta, p_X, p_{Y|X})$ for all $p_{Y|X}$. The relative minimax criterion based on $\alpha(p_{Y|X})$ of (6.1) yields a higher $E_i(R, p_X, p_{Y|X})$ for good channels and this is counterbalanced by a lower $E_\emptyset(R, p_X, p_{Y|X})$. Thus the primary advantage of the relative minimax approach is that $\alpha(p_{Y|X})$ can be chosen to more finely balance the error exponents across the range of channels of interest.

7 Compound Binary Symmetric Channel

We have evaluated the incorrect-message and erasure exponents of (5.24) for the compound BSC with crossover probability $\rho \in [\rho_{\min}, \rho_{\max}]$, where $0 < \rho_{\min} < \rho_{\max} \leq \frac{1}{2}$. The class \mathscr{W} may

be identified with the interval $[\rho_{\min}, \rho_{\max}]$, where ρ_{\min} and ρ_{\max} correspond to the cleanest and noisiest channels in \mathcal{W} , respectively. Denote by $h_2(\rho) \triangleq -\rho \log \rho - (1 - \rho) \log(1 - \rho)$ the binary entropy function, by $h_2^{-1}(\cdot)$ the inverse of that function over the range $[0, \frac{1}{2}]$, and by p_ρ the Bernoulli p.m.f. with parameter ρ .

Capacity of the BSC is given by $C(\rho) = 1 - h_2(\rho)$, and the sphere packing exponent by [1, p. 195]

$$\begin{aligned} E_{sp}(R, \rho) &= D(p_{\rho_R} \| p_\rho) \\ &= \rho_R \log \frac{\rho_R}{\rho} + (1 - \rho_R) \log \frac{1 - \rho_R}{1 - \rho}, \quad 0 \leq R \leq C(\rho), \end{aligned} \quad (7.1)$$

where $\rho_R = h_2^{-1}(1 - R) \geq \rho$. The optimal input distribution p_X is uniform at all rates and will be omitted from the list of arguments of the functions $E_{r,F}$, E_{sp} , and F_R below. The critical rate is

$$R_{cr}(\rho) = 1 - h_2 \left(\frac{1}{1 + \sqrt{1/\rho^2 - 1}} \right),$$

and $E_{sp}(0, \rho) = -\log \sqrt{4\rho(1 - \rho)}$.

The capacity and sphere-packing exponent for the compound BSC are respectively given by $C(\mathcal{W}) = C(\rho_{\max})$ and

$$E_{sp}(R, \mathcal{W}) = \min_{\rho_{\min} \leq \rho \leq \rho_{\max}} E_{sp}(R, \rho) = E_{sp}(R, \rho_{\max}). \quad (7.2)$$

For $R \geq C(\mathcal{W})$, the optimal F is the thresholding rule of (5.1), and (5.2) and (5.3) yield

$$E_i(R, \rho) = \Delta \quad \text{and} \quad E_\emptyset(R, \rho) = E_{sp}(R + \Delta, \rho).$$

In the remainder of this section we consider the case $R < C(\mathcal{W})$, in which case $E_{sp}(R, \mathcal{W}) > 0$.

Optimal F . For any $0 \leq \Delta \leq C(\rho) - R$, we have $\rho \leq \rho_{R+\Delta} \leq \rho_R \leq \frac{1}{2}$. From (7.1) we have

$$\begin{aligned} F_{R,\rho}(t) &= E_{sp}(R, \rho) - E_{sp}(R + t, \rho) \\ &= D(p_{\rho_R} \| p_\rho) - D(p_{\rho_{R+t}} \| p_\rho) \\ &= h_2(\rho_R) - h_2(\rho_{R+t}) + (\rho_R - \rho_{R+t}) \left(\log \frac{1}{\rho} - \log \frac{1}{1 - \rho} \right) \\ &= h_2(\rho_R) - h_2(\rho_{R+t}) + (\rho_R - \rho_{R+t}) \log \left(\frac{1}{\rho} - 1 \right), \quad t \geq 0, \end{aligned}$$

which is a decreasing function of ρ .

Evaluating the optimal F from (5.12), we have

$$\begin{aligned} F^{L*}(t) &= \Delta + F_{R,\mathcal{W}}(t) \\ &= \Delta + E_{sp}(R, \mathcal{W}) - E_{sp}(R + t, \mathcal{W}) \\ &= \Delta + E_{sp}(R, \rho_{\max}) - E_{sp}(R + t, \rho_{\max}) \\ &= \Delta + F_{R,\rho_{\max}}(t), \quad t \geq 0. \end{aligned}$$

Observe that the optimal F is determined by the noisiest channel (ρ_{\max}) and does not depend at all on ρ_{\min} .

This contrasts with the relative minimax criterion with $\alpha(p_{Y|X})$ of (6.1), where evaluation of the optimal F from (6.9) yields

$$\begin{aligned} F^{L*}(t) &= \Delta + \max_{\rho_{\min} \leq \rho \leq \rho_{\max}} F_{R,\rho}(t) \\ &= \Delta + F_{R,\rho_{\min}}(t), \quad t \geq 0 \end{aligned}$$

which is determined by the cleanest channel (ρ_{\min}) and does not depend on ρ_{\max} .

Optimal Error Exponents. The derivations below are simplified if instead of working with the crossover probability ρ , we use the following reparameterization:

$$\mu \triangleq \rho^{-1} - 1, \quad \mu_{\max} \triangleq \rho_{\min}^{-1} - 1, \quad \mu_{\min} \triangleq \rho_{\max}^{-1} - 1$$

and

$$\mu_R \triangleq \rho_R^{-1} - 1 = \frac{1}{h_2^{-1}(1-R)} - 1 \quad \Leftrightarrow \quad R = 1 - h_2\left(\frac{1}{1+\mu}\right)$$

where μ_R increases monotonically from 1 to ∞ as R increases from 0 to 1. With this notation, we have $\rho = \frac{1}{1+\mu}$ and $\mu_{\max} \geq \mu \geq \mu_{R+\Delta} \geq \mu_R \geq 1$. Also

$$\begin{aligned} \frac{dR}{d\rho_R} &= -\frac{dh_2(\rho_R)}{d\rho_R} = -\frac{\log \mu_R}{\ln 2} \\ \frac{dE_{sp}(R, \rho)}{d\rho_R} &= \frac{\log \mu - \log \mu_R}{\ln 2} \\ -E'_{sp}(R, \rho) &= \frac{dE_{sp}(R, \rho)/d\rho_R}{-dR/d\rho_R} = \frac{\log \mu}{\log \mu_R} - 1 = \frac{\log \mu/\mu_R}{\log \mu_R}. \end{aligned} \tag{7.3}$$

From (5.18) and (7.3), we obtain

$$-E'_{sp}(R, \mathcal{W}) = -\min_{\rho_{\min} \leq \rho \leq \rho_{\max}} E'_{sp}(R, \rho) = \frac{\log \mu_{\max}/\mu_R}{\log \mu_R}. \tag{7.4}$$

Observe that the minimizing ρ is ρ_{\min} , i.e., the cleanest channel in \mathcal{W} . In contrast, from (7.2), we have

$$-E'_{sp}(R, \mathcal{W}) = -E'_{sp}(R, \rho_{\max}) = \frac{\log \mu_{\min}/\mu_R}{\log \mu_R} < -E'_{sp}(R, \mathcal{W}) \tag{7.5}$$

which is determined by the noisiest channel (ρ_{\max}).

Conditions for universality. Next we evaluate $\overline{R}^{conj}(\mathcal{W})$ from (5.19). For a given ρ , the conjugate rate of R is obtained from (7.3):

$$\begin{aligned} -E'_{sp}(R, \rho) &= \frac{1}{-E'_{sp}(R^{conj}, \rho)} \\ \frac{\log(\mu/\mu_R)}{\log \mu_R} &= \frac{\log \mu_{R^{conj}}}{\log(\mu/\mu_{R^{conj}})} \end{aligned}$$

hence

$$\begin{aligned}
\mu_{R^{conj}} &= \frac{\mu}{\mu_R} \\
R^{conj}(\mu) &= 1 - h_2\left(\frac{1}{1 + \mu/\mu_R}\right) \\
\overline{R}^{conj}(\mathcal{W}) &= \max_{\mu_{\min} \leq \mu \leq \mu_{\max}} R^{conj}(\mu) \\
&= R^{conj}(\mu_{\max}).
\end{aligned}$$

From (5.10) and (7.2), we have

$$R^{conj}(\mathcal{W}) = R^{conj}(\mu_{\min}).$$

Analogously to (7.5), observe that both $\overline{R}^{conj}(\mathcal{W})$ and $R^{conj}(\mathcal{W})$ are determined by the cleanest and noisiest channels in \mathcal{W} , respectively.

We can now evaluate the conditions of Prop. 5.5, under which $F(t) = \Delta + \lambda|t|^+$ is universal (subject to conditions on Δ and λ). In (5.22), Δ must satisfy

$$\begin{aligned}
\left| \overline{R}^{conj}(\mathcal{W}) - R \right|^+ &\leq \Delta \leq C(\mathcal{W}) - R \\
\left| R^{conj}(\mu_{\max}) - R \right|^+ &\leq \Delta \leq C(\mu_{\min}) - R \\
\left| h_2\left(\frac{1}{1 + \mu_R}\right) - h_2\left(\frac{1}{1 + \mu_{\max}/\mu_R}\right) \right|^+ &\leq \Delta \leq h_2\left(\frac{1}{1 + \mu_R}\right) - h_2\left(\frac{1}{1 + \mu_{\min}}\right). \quad (7.6)
\end{aligned}$$

The left side is zero if $\mu_{\max} \leq \mu_R^2$. If $\mu_{\max} > \mu_R^2$, the argument of $|\cdot|^+$ is positive, and we need $\mu_{\max}/\mu_R < \mu_{\min}$ to ensure that the left side is lower than the right side. Hence there exists a nonempty range of values of Δ satisfying (7.6) if and only if

$$\mu_R \geq \min \left\{ \sqrt{\mu_{\max}}, \frac{\mu_{\max}}{\mu_{\min}} \right\}$$

which may also be written as

$$\mu_{\max} \leq \max\{\mu_R^2, \mu_R \mu_{\min}\}.$$

Next, substituting (7.4) into (5.23), we obtain the following condition for λ :

$$\frac{\log \mu_{\max}/\mu_R}{\log \mu_R} \leq \lambda \leq \frac{\log \mu_{R+\Delta}}{\log \mu_{\max}/\mu_{R+\Delta}}. \quad (7.7)$$

This equation has a solution if and only if the left side does not exceed the right side, i.e.,

$$\mu_{\max} \leq \mu_R \mu_{R+\Delta}, \quad (7.8)$$

or equivalently, $\rho_{\min} \geq (1 + \mu_R \mu_{R+\Delta})^{-1}$. Since μ_R is an increasing function of R , the larger the values of R and Δ , the lower the value of ρ_{\min} for which the universality property still holds.

If equality holds in (7.8), the only feasible value of λ is

$$\lambda = \frac{\log \mu_{R+\Delta}}{\log \mu_R} \geq 1. \quad (7.9)$$

This value of λ remains feasible if (7.8) holds with strict inequality.

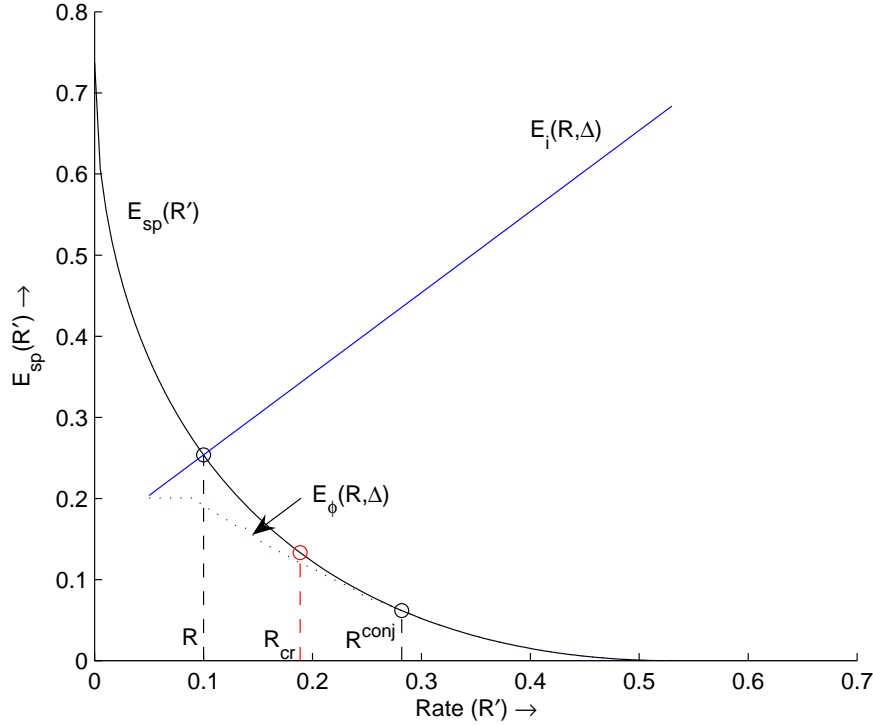


Figure 2: Erasure and incorrect-message exponents $E_0(R)$ and $E_i(R)$ for BSC with crossover probability $\rho = 0.1$ (capacity $C(\rho) \approx 0.53$) and rate $R = 0.1$.

8 Discussion

The \mathcal{F} -MMI decision rule of (3.1) is a generalization of Csiszár and Körner’s MMI decoder with erasure option. The weighting function F in (3.1) can be optimized in an asymptotic Neyman-Pearson sense given a compound class of channels \mathcal{W} . An explicit formula has been derived in terms of the sphere-packing exponent function for F that maximizes the erasure exponent subject to a constraint on the incorrect-message exponent. The optimal F is generally nonunique but agrees with existing designs in special cases of interest.

In particular, Corollary 5.2 shows that the simple thresholding rule $F(t) \equiv \Delta$ is optimal if $R \geq I(p_X, \mathcal{W})$, i.e., when the transmission rate cannot be reliably supported by the worst channel in \mathcal{W} . When $R < I(p_X, \mathcal{W})$, Prop. 5.5 shows that for small erasure exponents, our expressions for the optimal exponents coincide with those derived by Forney [3] for symmetric channels, where the same input distribution p_X is optimal at all rates. In this regime, Csiszár and Körner’s rule $F(t) = \Delta + \lambda|t|^+$ is also universal under some conditions on the parameter pair (Δ, λ) . It is also worth noting that while suboptimal, the design $F(t) = \Delta + t$ yields an empirical version of Forney’s simple decision rule (2.5).

Previous work [5] using a different universal decoder had shown that Forney’s exponents can be

matched in the special case where undetected-error and erasure exponents are equal (corresponding to $T = 0$ in Forney's rule (2.4)). Our results show that this property extends beyond this special case, albeit not everywhere.

Another analogy between Forney's suboptimal decision rule (2.5) and ours (3.1) is that the former is based on the two highest likelihood scores, and the latter is based on the two highest empirical mutual information scores. Our results imply that (2.5) is optimal (in terms of error exponents) in the special regime identified above.

The relative minimax criterion of Sec. 6 is attractive when the compound class \mathcal{W} is broad (or difficult to pick) as it allows finer tuning of the error exponents for different channels in \mathcal{W} . The class \mathcal{W} could conceivably be as large as $\mathcal{P}_{Y|X}$, the set of all DMC's.

Finally, we have extended our framework to decoding for compound multiple access channels. Those results will be presented elsewhere.

Acknowledgements. The author thanks Shankar Sadasivam for numerical evaluation of the error exponent formulas in Sec. 7, and Prof. Lizhong Zheng for helpful comments.

A Proof of Proposition 4.1

(i) is immediate from (4.3), restated below:

$$E_{r,F}(R, p_X, p_{Y|X}) = \min_{\tilde{p}_{Y|X}} [D(\tilde{p}_{Y|X} \| p_{Y|X} | p_X) + F(I(p_X, \tilde{p}_{Y|X}) - R)], \quad (\text{A.1})$$

and the fact that F is nondecreasing.

(ii) is immediate for the same reason as above.

(iii) Since the function $E_{sp}(R, p_X, p_{Y|X})$ is decreasing in R for $R \leq I(p_X, p_{Y|X})$, we have

$$E_{sp}(R, p_X, p_{Y|X}) = \min_{\tilde{p}_{Y|X} : I(p_X, \tilde{p}_{Y|X}) = R} D(\tilde{p}_{Y|X} \| p_{Y|X} | p_X), \quad \forall R \leq I(p_X, p_{Y|X}). \quad (\text{A.2})$$

Since F is nondecreasing, $\tilde{p}_{Y|X}$ that achieves the minimum in (A.1) must satisfy $I(p_X, \tilde{p}_{Y|X}) \leq I(p_X, p_{Y|X})$. Hence

$$\begin{aligned} E_{r,F}(R, p_X, p_{Y|X}) &= \min_{\tilde{p}_{Y|X} : I(p_X, \tilde{p}_{Y|X}) \leq I(p_X, p_{Y|X})} [D(\tilde{p}_{Y|X} \| p_{Y|X} | p_X) + F(I(p_X, \tilde{p}_{Y|X}) - R)] \\ &= \min_{R' \leq I(p_X, p_{Y|X})} \min_{\tilde{p}_{Y|X} : I(p_X, \tilde{p}_{Y|X}) = R'} [D(\tilde{p}_{Y|X} \| p_{Y|X} | p_X) + F(R' - R)] \\ &\stackrel{(a)}{=} \min_{R' \leq I(p_X, p_{Y|X})} [E_{sp}(R', p_X, p_{Y|X}) + F(R' - R)] \\ &\stackrel{(b)}{=} \min_{R'} [E_{sp}(R', p_X, p_{Y|X}) + F(R' - R)] \end{aligned}$$

where (a) is due to (A.2), and (b) holds because $E_{sp}(R', p_X, p_{Y|X}) = 0$ for $R' \geq I(p_X, p_{Y|X})$ and F is nondecreasing.

(iv) The claim follows directly from the definitions (4.2) and (4.4), taking minima over \mathcal{W} . \square

B Proof of Proposition 4.2

Given the p.m.f. p_X , choose any type $p_{\mathbf{x}}$ such that $\max_{x \in \mathcal{X}} |p_{\mathbf{x}}(x) - p_X(x)| \leq \frac{|\mathcal{X}|}{N}$.² Define

$$E_{r,F,N}(R, p_{\mathbf{x}}, p_{Y|X}) = \min_{p_{\mathbf{y}|\mathbf{x}}} [D(p_{\mathbf{y}|\mathbf{x}} \| p_{Y|X} | p_{\mathbf{x}}) + F(I(\mathbf{x}; \mathbf{y}) - R)] \quad (\text{B.1})$$

and

$$E_{sp,N}(R, p_{\mathbf{x}}, p_{Y|X}) = \min_{p_{\mathbf{y}|\mathbf{x}} : I(\mathbf{x}; \mathbf{y}) \leq R} D(p_{\mathbf{y}|\mathbf{x}} \| p_{Y|X} | p_{\mathbf{x}}) \quad (\text{B.2})$$

which differ from (4.3) and (4.1) in that the minimization is performed over conditional types instead of general conditional p.m.f.'s. We have

$$\begin{aligned} \lim_{N \rightarrow \infty} E_{r,F,N}(R, p_{\mathbf{x}}, p_{Y|X}) &= E_{r,F}(R, p_X, p_{Y|X}) \\ \lim_{N \rightarrow \infty} E_{sp,N}(R, p_{\mathbf{x}}, p_{Y|X}) &= E_{sp}(R, p_X, p_{Y|X}) \end{aligned} \quad (\text{B.3})$$

² For instance, truncate each $p_X(x)$ down to the nearest integer multiple of $1/N$ and add a/N to the smallest resulting value to obtain $p_{\mathbf{x}}(x)$, $x \in \mathcal{X}$, summing to one. a is an integer in the range $\{0, 1, \dots, |\mathcal{X}| - 1\}$.

by continuity of the divergence functional and of F in the region where the minimand $D + F$ is finite.

We will use the following two standard inequalities.

1) Given an arbitrary sequence \mathbf{y} , draw \mathbf{x}' independently of \mathbf{y} and uniformly over a fixed type class $T_{\mathbf{x}}$. Then [1]

$$Pr[T_{\mathbf{x}'|\mathbf{y}}] = \frac{|T_{\mathbf{x}'|\mathbf{y}}|}{|T_{\mathbf{x}}|} = \frac{|T_{\mathbf{x}'|\mathbf{y}}|}{|T_{\mathbf{x}'|}} \doteq 2^{-NI(\mathbf{x}';\mathbf{y})}.$$

Hence for any $0 \leq \nu \leq H(p_{\mathbf{x}})$,

$$\begin{aligned} Pr[I(\mathbf{x}';\mathbf{y}) \geq \nu] &= \sum_{T_{\mathbf{x}'|\mathbf{y}}} Pr[T_{\mathbf{x}'|\mathbf{y}}] \mathbb{1}_{\{I(\mathbf{x}';\mathbf{y}) \geq \nu\}} \\ &\doteq \sum_{T_{\mathbf{x}'|\mathbf{y}}} 2^{-NI(\mathbf{x}';\mathbf{y})} \mathbb{1}_{\{I(\mathbf{x}';\mathbf{y}) \geq \nu\}} \\ &\stackrel{(a)}{\doteq} \max_{T_{\mathbf{x}'|\mathbf{y}}} 2^{-NI(\mathbf{x}';\mathbf{y})} \mathbb{1}_{\{I(\mathbf{x}';\mathbf{y}) \geq \nu\}} \\ &\doteq 2^{-N\nu} \end{aligned} \tag{B.4}$$

where (a) holds because the number of types is polynomial in N . For $\nu > H(p_{\mathbf{x}})$ we have $Pr[I(\mathbf{x}';\mathbf{y}) \geq \nu] = 0$.

2) Given an arbitrary sequence \mathbf{x} , draw \mathbf{y} from the conditional p.m.f. $p_{Y|X}^N(\cdot|\mathbf{x})$. We have [1]

$$Pr[T_{\mathbf{y}|\mathbf{x}}] \doteq 2^{-ND(p_{\mathbf{y}|\mathbf{x}}\|p_{Y|X}|p_{\mathbf{x}})}. \tag{B.5}$$

Then for any $\nu > 0$,

$$\begin{aligned} Pr[I(\mathbf{x};\mathbf{y}) \leq \nu] &= \sum_{T_{\mathbf{y}|\mathbf{x}}} Pr[T_{\mathbf{y}|\mathbf{x}}] \mathbb{1}_{\{I(\mathbf{x};\mathbf{y}) \leq \nu\}} \\ &\doteq \sum_{p_{\mathbf{y}|\mathbf{x}}} 2^{-ND(p_{\mathbf{y}|\mathbf{x}}\|p_{Y|X}|p_{\mathbf{x}})} \mathbb{1}_{\{I(\mathbf{x};\mathbf{y}) \leq \nu\}} \\ &\doteq \max_{p_{\mathbf{y}|\mathbf{x}}} 2^{-ND(p_{\mathbf{y}|\mathbf{x}}\|p_{Y|X}|p_{\mathbf{x}})} \mathbb{1}_{\{I(\mathbf{x};\mathbf{y}) \leq \nu\}} \\ &= \max_{p_{\mathbf{y}|\mathbf{x}} : I(\mathbf{x};\mathbf{y}) \leq \nu} 2^{-ND(p_{\mathbf{y}|\mathbf{x}}\|p_{Y|X}|p_{\mathbf{x}})} \\ &= 2^{-NE_{sp,N}(\nu,p_{\mathbf{x}},p_{Y|X})}. \end{aligned} \tag{B.6}$$

Incorrect Messages. The codewords are drawn independently and uniformly from type class $T_{\mathbf{x}}$. Since the conditional error probability is independent of the transmitted message, assume without loss of generality that message $m = 1$ was transmitted. An incorrect codeword $\mathbf{x}(i)$ appears on the decoder's list if $i > 1$ and

$$I(\mathbf{x}(i);\mathbf{y}) \geq R + \max_{j \neq i} F(I(\mathbf{x}(j);\mathbf{y}) - R).$$

Let $\mathbf{x} = \mathbf{x}(1)$. To evaluate the expected number of incorrect codewords on the list, we first fix \mathbf{y} .

Given \mathbf{y} , define the i.i.d. random variables $Z_i = I(\mathbf{x}(i); \mathbf{y}) - R$ for $2 \leq i \leq 2^{NR}$. Also let $z_1 = I(\mathbf{x}; \mathbf{y}) - R$, which is a function of the joint type $p_{\mathbf{xy}}$. The expected number of incorrect codewords on the list depends on (\mathbf{x}, \mathbf{y}) only via their joint type and is given by

$$\begin{aligned}
\mathbb{E}[N_i | T_{\mathbf{xy}}] &= \sum_{i=2}^{2^{NR}} Pr \left[I(\mathbf{x}(i); \mathbf{y}) \geq R + \max_{j \notin \mathcal{M} \setminus \{i\}} F(I(\mathbf{x}(j); \mathbf{y}) - R) \right] \\
&= \sum_{i=2}^{2^{NR}} Pr \left[Z_i \geq \max_{j \notin \mathcal{M} \setminus \{i\}} F(Z_j) \right] \\
&\leq \sum_{i=2}^{2^{NR}} Pr[Z_i \geq F(z_1)] \\
&= (2^{NR} - 1) Pr[Z_2 \geq F(z_1)] \\
&\stackrel{(a)}{=} (2^{NR} - 1) Pr[I(\mathbf{x}'; \mathbf{y}) \geq R + F(I(\mathbf{x}; \mathbf{y}) - R)] \\
&\stackrel{(b)}{=} 2^{NR} 2^{-N[R + F(I(\mathbf{x}; \mathbf{y}) - R)]} \\
&= 2^{-NF(I(\mathbf{x}; \mathbf{y}) - R)}
\end{aligned} \tag{B.7}$$

where in (a), \mathbf{x}' is drawn independently of \mathbf{y} and uniformly over the type class $T_{\mathbf{x}}$; and (b) is obtained by application of (B.4).

Averaging over \mathbf{y} , we obtain

$$\begin{aligned}
\mathbb{E}[N_i | T_{\mathbf{x}}] &= \sum_{T_{\mathbf{y}|\mathbf{x}}} Pr[T_{\mathbf{y}|\mathbf{x}}] \mathbb{E}[N_i | T_{\mathbf{xy}}] \\
&\stackrel{(a)}{=} \max_{T_{\mathbf{y}|\mathbf{x}}} Pr[T_{\mathbf{y}|\mathbf{x}}] \mathbb{E}[N_i | T_{\mathbf{xy}}] \\
&\stackrel{(a)}{=} \max_{p_{\mathbf{y}|\mathbf{x}}} \exp_2 \{ -N[D(p_{\mathbf{y}|\mathbf{x}} \| p_{Y|X} | p_{\mathbf{x}}) + F(I(\mathbf{x}; \mathbf{y}) - R)] \} \\
&\stackrel{(b)}{=} \exp_2 \{ -NE_{r,F,N}(R, p_{\mathbf{x}}, p_{Y|X}) \} \\
&\stackrel{(c)}{=} \exp_2 \{ -NE_{r,F}(R, p_X, p_{Y|X}) \}
\end{aligned}$$

where (a), (b), (c) follow from (B.5), (B.1), (B.3), respectively. This proves (4.8).

Erasure. The decoder fails to return the transmitted codeword $\mathbf{x} = \mathbf{x}(1)$ if

$$I(\mathbf{x}; \mathbf{y}) \leq R + \max_{2 \leq i \leq 2^{NR}} F(I(\mathbf{x}(i); \mathbf{y}) - R). \tag{B.8}$$

Denote by $p_\emptyset(p_{\mathbf{x}})$ the probability of this event. The event is the disjoint union of events \mathcal{E}_1 and \mathcal{E}_2 below. The first one is

$$\mathcal{E}_1 : \quad I(\mathbf{x}; \mathbf{y}) \leq R + F(0). \tag{B.9}$$

Since F^{-1} is increasing, \mathcal{E}_1 is equivalent to $F^{-1}(I(\mathbf{x}; \mathbf{y}) - R) \leq 0$. The second event is

$$\begin{aligned}
\mathcal{E}_2 : \quad R + F(0) < I(\mathbf{x}; \mathbf{y}) &\leq R + \max_{2 \leq i \leq 2^{NR}} F(I(\mathbf{x}(i); \mathbf{y}) - R) \\
&= R + F \left(\max_{2 \leq i \leq 2^{NR}} I(\mathbf{x}(i); \mathbf{y}) - R \right)
\end{aligned}$$

where equality holds because F is nondecreasing. Thus \mathcal{E}_2 is equivalent to

$$\max_{2 \leq i \leq 2^{NR}} I(\mathbf{x}(i); \mathbf{y}) \geq R + F^{-1}(I(\mathbf{x}; \mathbf{y}) - R) > R. \quad (\text{B.10})$$

Applying (B.6), we have

$$\begin{aligned} Pr[\mathcal{E}_1] = p_\emptyset^{(1)}(p_{\mathbf{x}}) &\triangleq Pr[I(\mathbf{x}; \mathbf{y}) \leq R + F(0)] \\ &\doteq \exp_2 \{ -NE_{sp,N}(R + F(0), p_{\mathbf{x}}, p_{Y|X}) \}. \end{aligned} \quad (\text{B.11})$$

Clearly $p_\emptyset(p_{\mathbf{x}}) \sim p_\emptyset^{(1)}(p_{\mathbf{x}}) \sim 1$ if $R + F(0) \geq I(p_{\mathbf{x}}, p_{Y|X})$.

Next we have

$$\begin{aligned} Pr[\mathcal{E}_2] = p_\emptyset^{(2)}(T_{\mathbf{y}|\mathbf{x}}) &= Pr \left[\max_{2 \leq i \leq 2^{NR}} I(\mathbf{x}(i); \mathbf{y}) \geq R + F^{-1}(I(\mathbf{x}; \mathbf{y}) - R) \right] \\ &= Pr \left[\max_{2 \leq i \leq 2^{NR}} Z_i \geq F^{-1}(z_1) \right] \\ &\stackrel{(a)}{\doteq} 2^{-NF^{-1}(z_1)} \\ &= 2^{-NF^{-1}(I(\mathbf{x}; \mathbf{y}) - R)} \end{aligned}$$

where (a) follows by application of the union bound and (B.4).

Averaging over \mathbf{y} , we obtain

$$\begin{aligned} p_\emptyset^{(2)}(p_{\mathbf{x}}) &\triangleq \sum_{T_{\mathbf{y}|\mathbf{x}} : I(\mathbf{x}; \mathbf{y}) > R + F(0)} Pr[T_{\mathbf{y}|\mathbf{x}}] p_\emptyset^{(2)}(T_{\mathbf{y}|\mathbf{x}}) \\ &\doteq \max_{T_{\mathbf{y}|\mathbf{x}} : I(\mathbf{x}; \mathbf{y}) > R + F(0)} Pr[T_{\mathbf{y}|\mathbf{x}}] p_\emptyset^{(2)}(T_{\mathbf{y}|\mathbf{x}}) \\ &\doteq \max_{p_{\mathbf{y}|\mathbf{x}} : I(\mathbf{x}; \mathbf{y}) > R + F(0)} \exp_2 \{ -N[D(p_{\mathbf{y}|\mathbf{x}} \| p_{Y|X} | p_{\mathbf{x}}) + F^{-1}(I(\mathbf{x}; \mathbf{y}) - R)] \}. \end{aligned} \quad (\text{B.12})$$

For $R + F(0) \leq I(p_{\mathbf{x}}, p_{Y|X})$, we have

$$\max_{p_{\mathbf{y}|\mathbf{x}} : I(\mathbf{x}; \mathbf{y}) \leq R + F(0)} \exp_2 \{ -ND(p_{\mathbf{y}|\mathbf{x}} \| p_{Y|X} | p_{\mathbf{x}}) \} \doteq \max_{p_{\mathbf{y}|\mathbf{x}} : I(\mathbf{x}; \mathbf{y}) = R + F(0)} \exp_2 \{ -ND(p_{\mathbf{y}|\mathbf{x}} \| p_{Y|X} | p_{\mathbf{x}}) \},$$

hence (B.12) may be written more simply as

$$\begin{aligned} p_\emptyset^{(2)}(p_{\mathbf{x}}) &\doteq \max_{p_{\mathbf{y}|\mathbf{x}}} \exp_2 \{ -N[D(p_{\mathbf{y}|\mathbf{x}} \| p_{Y|X} | p_{\mathbf{x}}) + |F^{-1}(I(\mathbf{x}; \mathbf{y}) - R)|^+] \} \\ &= \exp_2 \{ -NE_{r, |F^{-1}|^+, N}(R, p_{\mathbf{x}}, p_{Y|X}) \}. \end{aligned} \quad (\text{B.13})$$

Since \mathcal{E}_1 and \mathcal{E}_2 are disjoint events, we obtain

$$\begin{aligned} p_\emptyset(p_{\mathbf{x}}) &= p_\emptyset^{(1)}(p_{\mathbf{x}}) + p_\emptyset^{(2)}(p_{\mathbf{x}}) \\ &\doteq \exp_2 \{ -N \min \{ E_{r, |F^{-1}|^+, N}(R, p_{\mathbf{x}}, p_{Y|X}), E_{sp,N}(R + F(0), p_{\mathbf{x}}, p_{Y|X}) \} \} \\ &\doteq \exp_2 \{ -N \min \{ E_{r, |F^{-1}|^+}(R, p_X, p_{Y|X}), E_{sp}(R + F(0), p_X, p_{Y|X}) \} \} \end{aligned}$$

where the last line is due to (B.3).

The function $F^{-1}(t)$ has a zero-crossing at $t = F(0)$. Applying (4.7), we obtain

$$\begin{aligned} E_{r,|F^{-1}|+}(R, p_X, p_{Y|X}) &= \min_{R'} [E_{sp}(R', p_X, p_{Y|X}) + |F^{-1}(R' - R)|^+] \\ &\leq E_{sp}(R + F(0), p_X, p_{Y|X}) + 0 \end{aligned}$$

hence

$$p_\emptyset(p_{\mathbf{x}}) \doteq \exp_2\{-NE_{r,|F^{-1}|+}(R, p_X, p_{Y|X})\}.$$

This proves (4.9). \square

C Proof of Proposition 5.1

We first prove (5.12). Recall (4.7) and (4.6), restated here for convenience:

$$\begin{aligned} E_{r,F}(R, p_X, \mathcal{W}) &= \min_{R'} [E_{sp}(R', p_X, \mathcal{W}) + F(R' - R)], \\ E_{sp}(R, p_X, \mathcal{W}) &\equiv E_{sp}(R', p_X, \mathcal{W}) + F_{R,p_X,\mathcal{W}}(R' - R), \quad \forall R'. \end{aligned}$$

Hence

$$E_{r,F_{R,p_X,\mathcal{W}}}(R, p_X, \mathcal{W}) = E_{sp}(R, p_X, \mathcal{W}). \quad (\text{C.1})$$

Case I: $E_{sp}(R, p_X, \mathcal{W}) = \alpha$, i.e., from (5.9), we have $\Delta = 0$. The feasible set $\mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$ defined in (5.6) takes the form $\{F : E_{r,F}(R, p_X, \mathcal{W}) \geq E_{sp}(R, p_X, \mathcal{W})\}$. Owing to (C.1) and the monotonicity property of Prop. 4.1(ii), an equivalent representation of $\mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$ is $\{F : F \succeq F_{R,p_X,\mathcal{W}}\}$. As indicated below the statement of Prop. 5.1, this implies $F_{R,p_X,\mathcal{W}}$ achieves the supremum in (5.5).

Case II: $E_{sp}(R, p_X, \mathcal{W}) \neq \alpha$. The derivation parallels that of Case I. An equivalent representation of the constrained set $\mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$ in (5.6) is $\{F : F \succeq F^{L*}\}$, where

$$\begin{aligned} F^{L*}(t) &= F_{R,p_X,\mathcal{W}}(t) + \alpha - E_{sp}(R, p_X, \mathcal{W}) \\ &= F_{R,p_X,\mathcal{W}}(t) + \Delta \\ &= \alpha - E_{sp}(R + t, p_X, \mathcal{W}). \end{aligned}$$

Hence F^{L*} achieves the maximum in (5.5).

To prove (5.13), we simply observe that if $F^{L*} \preceq F$ for all $F \in \mathcal{F}^L(R, p_X, \mathcal{W}, \alpha)$, then $F^* \preceq F$ for all $F \in \mathcal{F}(R, p_X, \mathcal{W}, \alpha)$, where $F^*(t) = \max(t, F^{L*}(t))$. \square

D Proof of Proposition 5.3

We have

$$\begin{aligned} E_\emptyset^L(R, p_X, \mathcal{W}, \alpha) &\stackrel{(a)}{=} E_{r,|(F^{L*})^{-1}|+}(R, p_X, \mathcal{W}) \\ &\stackrel{(b)}{=} \min_{R'} [E_{sp}(R', p_X, \mathcal{W}) + |(F^{L*})^{-1}|^+(R' - R)] \\ &\stackrel{(c)}{=} \min_{R' \geq R + t_{(F^{L*})^{-1}}} [E_{sp}(R', p_X, \mathcal{W}) + (F^{L*})^{-1}(R' - R)] \end{aligned} \quad (\text{D.1})$$

where equality (a) results from Props. 4.2 and 5.1, (b) follows from (4.7), and (c) from the definition (3.12) and the fact that the function $E_{sp}(R', p_X, \mathcal{W})$ is nonincreasing in R' . From (5.12) and property (P3) in Sec. 3 we obtain the inverse function

$$(F^{L*})^{-1}(t) = F_{R, p_X, \mathcal{W}}^{-1}(t - \Delta) \quad (\text{D.2})$$

where $F_{R, p_X, \mathcal{W}}(t)$ is given in (4.5). Hence $t_{(F^{L*})^{-1}} = \Delta$ and

$$E_{\emptyset}^L(R, p_X, \mathcal{W}, \alpha) = \min_{R' \geq R + \Delta} \underbrace{[E_{sp}(R', p_X, \mathcal{W}) + F_{R, p_X, \mathcal{W}}^{-1}(R' - R - \Delta)]}_{h(R')}. \quad (\text{D.3})$$

By assumption $E_{sp}(R, p_X, \mathcal{W})$ is convex in R , and therefore $F_{R, p_X, \mathcal{W}}(t)$ is concave. By application of Property (P5) in Sec. 3, the function $F_{R, p_X, \mathcal{W}}^{-1}$ is convex, and thus so is $h(R')$ in (D.3). The derivatives of $F_{R, p_X, \mathcal{W}}^{-1}$ and h are respectively given by

$$(F_{R, p_X, \mathcal{W}}^{-1})'(t) = \frac{1}{F'_{R, p_X, \mathcal{W}}(t)} = \frac{1}{-E'_{sp}(R + t, p_X, \mathcal{W})} \quad (\text{D.4})$$

and

$$h'(R') = E'_{sp}(R', p_X, \mathcal{W}) + \frac{1}{-E'_{sp}(R' - \Delta, p_X, \mathcal{W})}.$$

By Prop. 5.1 and the definition of Δ in (5.9), we have $E_i(R, p_X, \mathcal{W}) = \alpha = E_{sp}(R, p_X, \mathcal{W}) + \Delta$. Next we prove the statements (i)—(iv).

(i) $\max(0, R^{conj}(p_X, \mathcal{W}) - R) \leq \Delta \leq I(p_X, \mathcal{W}) - R$.

This case is illustrated in Fig. 3. We have

$$R + \Delta \geq \max(R, R^{conj}(p_X, \mathcal{W})) \geq R_{cr}(p_X, \mathcal{W}). \quad (\text{D.5})$$

Hence

$$\begin{aligned} h'(R + \Delta) &= E'_{sp}(R + \Delta, p_X, \mathcal{W}) + \frac{1}{-E'_{sp}(R, p_X, \mathcal{W})} \\ &\geq \begin{cases} E'_{sp}(R, p_X, \mathcal{W}) + \frac{1}{-E'_{sp}(R, p_X, \mathcal{W})} \geq 0 & : \text{ if } R \geq R_{cr}(p_X, \mathcal{W}) \\ E'_{sp}(R^{conj}(p_X, \mathcal{W}), p_X, \mathcal{W}) + \frac{1}{-E'_{sp}(R, p_X, \mathcal{W})} = 0 & : \text{ if } R \leq R_{cr}(p_X, \mathcal{W}) \end{cases} \\ &\geq 0. \end{aligned} \quad (\text{D.6})$$

By convexity of $h(\cdot)$ this implies that $R + \Delta$ minimizes $h(R')$ over $R' \geq R + \Delta$, and so

$$E_{\emptyset}^L(R, p_X, \mathcal{W}, \alpha) = h(R + \Delta) = E_{sp}(R + \Delta, p_X, \mathcal{W}).$$

(ii). Due to (D.5), we have either $R \geq R_{cr}(p_X, \mathcal{W})$ or $R + \Delta \geq R^{conj}(p_X, \mathcal{W}) \geq R$. In both cases,

$$-E'_{sp}(R, p_X, \mathcal{W}) \leq \frac{1}{-E'_{sp}(R + \Delta, p_X, \mathcal{W})}.$$

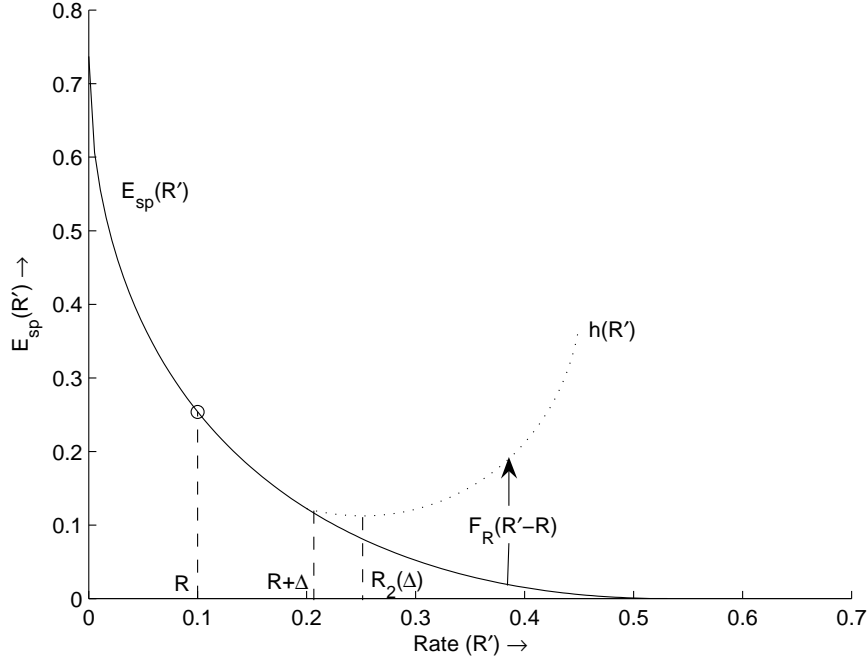


Figure 3: Construction and minimization of $h(R')$ for Case (i).

Let $F(t) = \Delta + \lambda|t|^+$ where λ is sandwiched by the left and right sides of the above inequality. We have $t_F = 0$ and $F'(t) = \lambda \mathbb{1}_{\{t \geq 0\}}$. The inverse function is $F^{-1}(t) = \frac{1}{\lambda}(t - \Delta)$ for $t \geq \Delta$. Hence $(F^{-1})'(t) = \frac{1}{\lambda} \mathbb{1}_{\{t \geq \Delta\}}$ and $t_{F^{-1}} = \Delta$. Substituting F and F^{-1} into (4.7), we obtain

$$E_{r,F}(R, p_X, \mathscr{W}) = \min_{R' \geq R} [E_{sp}(R', p_X, \mathscr{W}) + \Delta + \lambda(R' - R)],$$

$$E_{r,|F^{-1}|+}(R, p_X, \mathscr{W}) = \min_{R' \geq R+\Delta} \left[E_{sp}(R', p_X, \mathscr{W}) + \frac{1}{\lambda}(R' - R - \Delta) \right].$$

Taking derivatives of the bracketed terms with respect to R' and recalling that

$$\lambda \geq -E'_{sp}(R, p_X, \mathscr{W}), \quad \frac{1}{\lambda} \geq -E'_{sp}(R + \Delta, p_X, \mathscr{W}),$$

we observe that these derivatives are nonnegative. Since $E_{sp}(\cdot, p_X, \mathscr{W})$ is convex, the minima are achieved at R and $R + \Delta$ respectively.

The resulting exponents are $E_{sp}(R, p_X, \mathscr{W}) + \Delta$ and $E_{sp}(R + \Delta, p_X, \mathscr{W})$ which coincide with the optimal exponents of (5.14).

(iii). $R \leq R_{cr}(p_X, \mathscr{W})$ and $0 \leq \Delta \leq R^{conj}(p_X, \mathscr{W}) - R$.

This case is illustrated in Fig. 4 in the case $\Delta = 0$. From (D.6), we have $h'(R') = 0$ if and only if R' and $R' - \Delta$ are conjugate rates. In this case, using the above assumption on Δ , we have

$$R \leq R' - \Delta = R_1(\Delta) \leq R_{cr}(p_X, \mathscr{W}) \leq R' = R_2(\Delta) \leq R^{conj}(p_X, \mathscr{W}). \quad (\text{D.7})$$

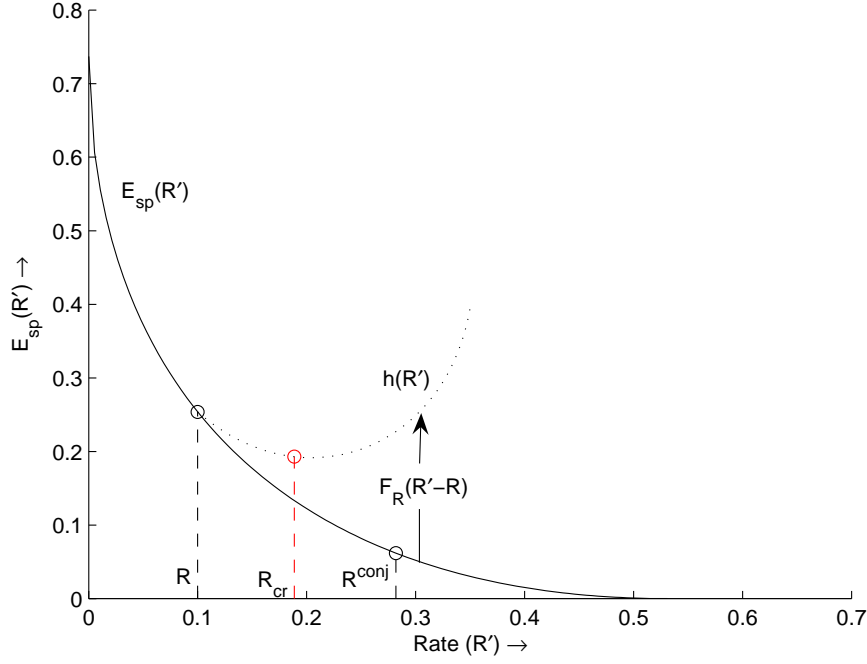


Figure 4: Construction and minimization of $h(R')$ for Case (ii), with $\Delta = 0$.

Hence $R_2(\Delta) = R' \geq R + \Delta$ is feasible for (D.3) and minimizes $h(\cdot)$. Substituting R' back into (D.3), we obtain

$$E_{\emptyset}^L(R, p_X, \mathcal{W}, \alpha) = h(R_2(\Delta)) = E_{sp}(R_2(\Delta), p_X, \mathcal{W}) + F_{R, p_X, \mathcal{W}}^{-1}(R_1(\Delta) - R)$$

which establishes (5.16).

(iv). $R \leq R_{cr}(p_X, \mathcal{W})$ and $R_{\infty}(p_X, \mathcal{W}) - R \leq \Delta \leq 0$.

Again we have $h'(R') = 0$ if and only if R' and $R' - \Delta$ are conjugate rates. Then, using the above assumption on Δ , we have

$$R \leq R' = R_1(\Delta) \leq R_{cr}(p_X, \mathcal{W}) \leq R' - \Delta = R_2(\Delta). \quad (\text{D.8})$$

Hence $R_1(\Delta) = R' \geq R + \Delta$ is feasible for (D.3) and minimizes $h(\cdot)$. Substituting R' back into (D.3), we obtain

$$E_{\emptyset}^L(R, p_X, \mathcal{W}, \alpha) = h(R_1(\Delta)) = E_{sp}(R_1(\Delta), p_X, \mathcal{W}) + F_{R, p_X, \mathcal{W}}^{-1}(R_2(\Delta) - R)$$

which establishes (5.17). □

E Proof of Lemma 5.4

First we prove (5.20). For any $R_0 < R_1$, we have

$$\begin{aligned}
\int_{R_0}^{R_1} \underline{E}'_{sp}(R, p_X, \mathcal{W}) dR &\stackrel{(a)}{=} \int_{R_0}^{R_1} \min_{p_{Y|X} \in \mathcal{W}} E'_{sp}(R, p_X, p_{Y|X}) dR \\
&\leq \min_{p_{Y|X} \in \mathcal{W}} \int_{R_0}^{R_1} E'_{sp}(R, p_X, p_{Y|X}) dR \\
&= \min_{p_{Y|X} \in \mathcal{W}} [E_{sp}(R_1, p_X, p_{Y|X}) - E_{sp}(R_0, p_X, p_{Y|X})] \\
&\stackrel{(b)}{\leq} E_{sp}(R_1, p_X, p_{Y|X}^*) - E_{sp}(R_0, p_X, p_{Y|X}^*) \\
&= \min_{p_{Y|X} \in \mathcal{W}} E_{sp}(R_1, p_X, p_{Y|X}) - E_{sp}(R_0, p_X, p_{Y|X}^*) \\
&\leq \min_{p_{Y|X} \in \mathcal{W}} E_{sp}(R_1, p_X, p_{Y|X}) - \min_{p_{Y|X} \in \mathcal{W}} E_{sp}(R_0, p_X, p_{Y|X}) \\
&= E_{sp}(R_1, p_X, \mathcal{W}) - E_{sp}(R_0, p_X, \mathcal{W}) \\
&= \int_{R_0}^{R_1} E'_{sp}(R, p_X, \mathcal{W}) dR \tag{E.1}
\end{aligned}$$

where (a) follows from the definition of \underline{E}'_{sp} in (5.18), and we choose $p_{Y|X}^*$ in inequality (b) as the minimizer of $E_{sp}(R_1, p_X, \cdot)$ over \mathcal{W} . Since (E.1) holds for all $R_0 < R_1$, we must have inequality between the integrands in the left and right sides: $\underline{E}'_{sp}(R, p_X, \mathcal{W}) \leq E'_{sp}(R, p_X, \mathcal{W})$. Moreover, the three inequalities used to derive (E.1) hold with equality if the same $p_{Y|X}^*$ minimizes $E'_{sp}(R, p_X, \cdot)$ at all rates, and the same $p_{Y|X}^*$ minimizes $E_{sp}(R, p_X, \cdot)$ at all rates. We need not (and generally do not) have $p_{Y|X}^{**} = p_{Y|X}^*$.³

Next we prove (5.21). By definition of $R^{conj}(p_X, \mathcal{W})$, we have

$$\begin{aligned}
-E'_{sp}(R^{conj}(p_X, \mathcal{W}), p_X, \mathcal{W}) &= \frac{1}{-E'_{sp}(R, p_X, \mathcal{W})} \\
&\stackrel{(a)}{\geq} \frac{1}{-\underline{E}'_{sp}(R, p_X, \mathcal{W})} = \min_{p_{Y|X} \in \mathcal{W}} \frac{1}{-E'_{sp}(R, p_X, p_{Y|X})} = \min_{p_{Y|X} \in \mathcal{W}} [-E'_{sp}(R^{conj}(p_X, p_{Y|X}), p_X, p_{Y|X})] \\
&\stackrel{(b)}{\geq} \min_{p_{Y|X} \in \mathcal{W}} [-E'_{sp}(\bar{R}^{conj}(p_X, \mathcal{W}), p_X, p_{Y|X})] \\
&\stackrel{(c)}{=} -\underline{E}'_{sp}(\bar{R}^{conj}(p_X, \mathcal{W}), p_X, \mathcal{W}) \\
&\stackrel{(d)}{\geq} -E'_{sp}(\bar{R}^{conj}(p_X, \mathcal{W}), p_X, \mathcal{W}).
\end{aligned}$$

where (a) and (d) are due to (5.20), (b) to the definition of $\bar{R}^{conj}(p_X, \mathcal{W})$ in (5.19) and the fact that $-E'_{sp}(R, p_X, p_{Y|X})$ is a decreasing function of R , and (c) from (5.18). Since $-E'_{sp}(R, p_X, \mathcal{W})$ is also a decreasing function of R , we must have $R^{conj}(p_X, \mathcal{W}) \leq \bar{R}^{conj}(p_X, \mathcal{W})$. Moreover, the conditions for equality are the same as those for equality in (5.20). \square

³ While $p_{Y|X}^*$ is the noisiest channel in \mathcal{W} , $p_{Y|X}^{**}$ may be the cleanest channel in \mathcal{W} , as in the BSC example of Sec. 7.

References

- [1] I. Csiszár and J. Körner, *Information Theory: Coding Theory for Discrete Memoryless Systems*, Academic Press, NY, 1981.
- [2] A. Lapidoth and P. Narayan, “Reliable Communication Under Channel Uncertainty,” *IEEE Trans. Information Theory*, Vol. 44, No. 6, pp. 2148—2177, Oct. 1998.
- [3] G. D. Forney, Jr., “Exponential Error Bounds for Erasure, List, and Decision Feedback Schemes,” *IEEE Trans. Information Theory*, Vol. 14, No. 2, pp. 206—220, 1968.
- [4] I. E. Telatar and R. G. Gallager, “New Exponential Upper Bounds to Error and Erasure Probabilities,” *Proc. ISIT’94*, p. 379, Trondheim, Norway, June 1994.
- [5] N. Merhav and M. Feder, “Minimax Universal Decoding with Erasure Option,” *IEEE Trans. Information Theory*, Vol. 53, No. 5, pp. 1664—1675, May 2007.
- [6] W. Hoeffding, “Asymptotically Optimal Tests for Multinomial Distributions,” *Ann. Math. Stat.*, Vol. 36, No. 2, pp. 369—400, 1965.
- [7] G. Tusnády, “On Asymptotically Optimal Tests,” *Annals of Statistics*, Vol. 5, No. 2, pp. 385—393, 1977.
- [8] O. Zeitouni and M. Gutman, “On Universal Hypotheses Testing via Large Deviations,” *IEEE Trans. Information Theory*, Vol. 37, No. 2, pp. 285—290, 1991.
- [9] I. Csiszár, *personal communication*, Aug. 2007.
- [10] M. Feder and N. Merhav, “Universal Composite Hypothesis Testing: A Competitive Minimax Approach,” *IEEE Trans. Information Theory*, Vol. 48, No. 6, pp. 1504—1517, June 2002.